DOCUMENT RESUME

ED 389 742 TM 024 371

AUTHOR Lawrence, Ida M.; And Others
TITLE DIF Data on Free-Response SAT I Mathematical Items.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-22
PUB DATE Aug 95
NOTE 77p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Black Students; Comparative Analysis; *Item Bias; *Mathematics Tests; Multiple Choice Tests; Pretests Posttests; *Racial Differences; *Sex Differences; Test Format; *Test Items; White Students
IDENTIFIERS *Free Response Test Items; *Scholastic Aptitude Test

ABSTRACT

This research summarizes differential item functioning (DIF) results for student produced response (SPR) items, a nonmultiple-choice mathematical item type in the Scholastic Aptitude Test I (SAT I). DIF data from 4 field trial pretest administrations (620 SPR items) and 10 final forms (100 SPR items with samples ranging from about 58,000 to over 250,000) were examined. For the male/female comparison, the SPR item type exhibited negligible DIF. The male/female DIF results for the SPR items resembled DIF results for the multiple-choice items in the SAT. For the African American/White comparison, SPR items exhibited more DIF than multiple-choice items. (Moderate or sizable DIF on multiple choice items in the old and new SAT-M is rare.) In particular, when African American test-takers were compared to matched groups of White test-takers, they performed less well than White test-takers on the SPR items. Although the level of African American/White DIF on SPR items was elevated, relative to the multiple-choice items, none of the SPR items in any of the final forms were flagged for exhibiting extreme DIF. The relationship between DIF and discrimination is discussed as a possible explanation for the elevated level of DIF on the SPR items. An appendix contains eight scatter plots. (Contains 6 tables, 14 figures, and 18 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 742

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DIF DATA ON FREE-RESPONSE SAT I MATHEMATICAL ITEMS

Ida M. Lawrence
C. Felicia Lyu
Miriam D. Feigenbaum

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
August 1995

TM024371

DIF Data on Free-Response SAT I Mathematical Items[1]

Ida M. Lawrence[2]

C. Felicia Lyu

Miriam D. Feigenbaum

Educational Testing Service

[1]Funding for this research came from ETS and the College Board.

[2]The authors thank Doris Larsen for help producing the DIF data, and Brent Bridgeman, Gary Marco, and Nancy Wright for reviewing earlier versions of this paper. In particular, the authors thank Rebecca Zwick for her reviews of several drafts of this paper and suggested revisions.

Copyright © 1995. Educational Testing Service. All rights reserved.

Abstract

The present research summarized Differential Item Functioning (DIF) results for Student Produced Response (SPR) items, a non-multiple choice mathematical item type in the SAT I. DIF data from four field trial pretest administrations (620 SPR items) and ten final forms (100 SPR items) were examined. For the Male/Female comparison, the SPR item type exhibits negligible DIF. The Male/Female DIF results for the SPR items resemble DIF results for the multiple choice items in the SAT. For the African American/White comparison, SPR items exhibit more DIF than the multiple choice items. (Moderate or sizable DIF on multiple choice items in the old or new SAT-M is rare.) In particular, when African American test takers are compared to matched groups of White test takers, they tend to perform less well than White test takers on the SPR items. Although the level of African American/White DIF on SPR items is elevated, relative to the multiple choice items, none of the SPR items in any of the final forms were flagged for exhibiting extreme DIF. The relationship between DIF and discrimination is discussed as a possible explanation for the elevated level of DIF on the SPR items.

DIF Data on Free-Response SAT I Mathematical Items

Ida M. Lawrence, C. Felicia Lyu and Miriam D. Feigenbaum

Educational Testing Service

The purpose of this paper is to evaluate differential item functioning (DIF) between female and male test takers and between African American and White test takers on the Student Produced Response (SPR) item type in the mathematical portion of the new SAT I. Along with other changes in the test, the SPR item type, which is a free response item, was introduced in the SAT I beginning with the March 1994 administration.

The mathematical portion of the SAT I measures developed mathematical reasoning abilities related to successful performance in college, and is intended to supplement the secondary school record and other information about the student. In terms of content, the questions in the mathematical test are distributed as follows: arithmetic reasoning (18-19 questions), algebraic reasoning (17 questions), geometric reasoning (16-17 questions) and miscellaneous reasoning (7-9 questions that focus on probability, data interpretation, or logical analysis).

The mathematical portion of the new SAT I is administered in two 30-minute sections plus one 15-minute section. Calculator use is permitted, but not required. The test is composed of 60 questions. Ten questions, the SPRs, require students to produce their own response and enter it in special grids on the answer sheet. The remaining 50 items in the test are multiple choice items. Each test contains 35 five-choice questions and 15 four-choice questions. The five-choice questions present a problem and ask the test taker to choose an answer from among the five options. The four-choice

Quantitative Comparison (QC) questions present two quantities (A and B) and the test taker must compare them and choose from four options: (1) quantity A is greater than quantity B, (2) quantity B is greater than quantity A, (3) the quantities are equal, or (4) the test taker cannot determine which quantity is greater from the information given.

The SPR items cover content that is similar to the other types of math questions in the test. However, the use of a grid for answering a question provides a different problem-solving situation because the test taker must answer the question without being influenced by various alternatives. In particular, the grid provides no feedback regarding the possible correctness of the answer. According to Braswell (1991), the SPR item type has several advantages. In particular, the SPR item type provides a more natural problem-solving situation, and allows test takers to answer the question in a way that is consistent with their solution. Also, the SPR item type discourages test-taking strategies such as plugging in each answer-choice. Finally, it is almost impossible to guess the correct answer. For more detail about the rationale for including the SPR item type in the new SAT I, see Braswell (1991) and Fremer (1993).

On the SAT I, questions answered correctly receive one point and omitted answers receive no points. A fraction of a point is subtracted for each multiple-choice question answered incorrectly (1/4 point for five-choice questions and 1/3 point for four-choice questions). No points are deducted for incorrect answers to the SPR questions.

Method

Samples and Instrument

Data for this study came from two sources:

1. SPR items administered in field trial pretests
2. SPR items administered in SAT I final forms

The configurations of the pretests and final forms are displayed in Table 1.

SPR Items Administered in Field Trial Pretests

Sixty-two different pretests containing SPR items were administered to classes in high schools between 1990 and 1992. The high schools volunteered to participate in four different field trials. This pretesting effort was part of a series of field trials carried out to obtain pretest statistics on new SAT I item types that could not be pretested in the variable section of the old SAT. Each pretest began with 25 four-choice QC items followed by 10 SPR items. Testing time for each pretest was 35 minutes. Students were permitted to use a calculator for these pretests. Table 2 presents ranges of sample sizes for the Male, Female, African American, and White test takers used in the DIF analyses of items administered in the field trial pretests. The pretest DIF analyses are based on non-representative samples from high schools who volunteered to participate in the field trials. For some of the pretest administrations, an effort was made to recruit high schools with high proportions of African American students.

SPR items administered in SAT I final forms

Of the 620 SPR items pretested in the field trials, 100 were used (some with minimal revision) in ten SAT I forms administered in 1994 at national test administrations. Table 3 presents sample sizes for each of the gender and ethnic groups used in the DIF analyses of SPR items administered in the final forms.

Differential Item Functioning (DIF)

The purpose of DIF assessment is to identify differences in item functioning for two comparable groups; the two groups are referred to as the reference group and the focal group. Items that function differently for matched reference groups and focal groups are identified as exhibiting DIF. For the DIF methodology used in this research, the focal group and the reference group are matched on a measure of the construct of interest. The total score on the test usually serves as the matching variable.

The procedure used to detect DIF was the Mantel-Haenszel (1959) procedure, adapted by Holland and Thayer (1988) for DIF assessment at ETS. The DIF statistic estimated with this procedure is an estimate of the odds ratio, under the assumption that the odds ratio is constant. the DIF statistic is defined as:

$$MH = [\Sigma_m R_{rm} W_{fm} / N_{tm}] / [\Sigma_m R_{fm} W_{rm} / N_{tm}]$$

where,

R=number right

W=number wrong

m=raw score on the matching criterion

fm=focal group at ability m

rm=reference group at ability m

tm=total group at ability m

The MH statistic is then transformed into a difference in deltas, one of the metrics used for reporting item difficulty to test developers at Educational Testing Service (ETS).

$$\text{MH D-DIF} = -2.35 \ln[\text{MH}]$$

The MH D-DIF statistic ranges from negative infinity to infinity with a value of 0 indicating no DIF. The magnitude of the MH D-DIF value and a significance test is used (at Educational Testing Service) to evaluate the presence or absence of DIF.

- MH D-DIF values are considered negligible if they are less than 1.0 in absolute value or are not statistically different from zero at the .05 significance level.

- MH D-DIF values between 1.0 and 1.5 in absolute value, or greater than 1.5 in absolute value and not significantly different from 1.0 at the .05 significance level, are considered moderate.

- MH D-DIF values are considered sizable if they exceed 1.5 in absolute value and are statistically different from 1.0 at the .05 significance level.

Items exhibiting sizable DIF are not to be included in a final form unless they are needed to meet specific statistical or content specifications. Items exhibiting moderate DIF are usually not selected for inclusion in a test unless the number of items exhibiting negligible DIF, is insufficient to meet statistical or content specifications. Note that average DIF is constrained to be

approximately zero across test items when an internal matching criterion is used. For a complete description of the Mantel-Haenszel procedure for DIF detection, see Dorans and Holland (1993).

Reference and Focal Groups

The same focal and reference groups were used in all of the DIF analyses in this study. Females defined the focal group and males defined the reference group for the Male/Female comparison. African Americans defined the focal group and Whites defined the reference group for the White/African American comparison. Samples were restricted to college bound juniors and seniors who identified themselves as having English as one of their first languages. In 1993, the mean mathematical scores on SAT (college-bound seniors) for females, males, African Americans, and Whites were 460, 501, 388, and 495, respectively (College Board, 1994).

Matching Criteria

The matching criterion for the DIF analyses of the SPR pretest items was the raw score on the 25 QC items (an "external" criterion) plus the studied item. As suggested by Holland and Thayer (1988), the studied item was included as part of the criterion when DIF analysis was performed on the item. The matching criterion for the primary DIF analyses of the SPR items in the final forms was the raw score on the total 60-item test (an "internal" criterion). In addition, supplemental DIF analyses on 40 SPR items (ten SPR items from four of the final forms) were carried out using a subset of items as the matching criterion - the 50 multiple choice items in each final form (an "external" criterion). Thus, for the supplemental analyses, the matching criterion did not contain any SPR items. Table 4 shows reliability estimates

(KR-20 with an adjustment for formula scoring) for the three possible matching variables.

As was done in the pretest analyses, the studied item was included as part of the criterion when DIF analysis was performed on the item. Using a matching criterion that contains only multiple choice items is analogous to the kind of external matching criterion (25 four-choice quantitative comparison items) used in the DIF analyses on the pretest items, although they included both four- and five-choice items.

Results

Frequency Distributions (Pretest Items)

Male/Female comparison.

Figure 1 shows frequency distributions of the MH D-DIF values from the Male/Female comparison for 620 SPR items pretested in four field trial administrations. Consistent across the four field trial administrations, Figure 1 indicates that the majority of MH D-DIF values fall in the interval between -1.0 and +1.0, indicating that DIF for females on the SPR pretest items tends to be negligible. A frequency distribution of the DIF values across the four administrations combined is displayed in Figure 2. The average MH D-DIF value for the 620 SPR items is -.11 (SD = .70).

White/African American comparison.

Figure 3 shows field trial administration specific frequency distributions of the MH D-DIF values from the White/African American comparison for 617 of the 620 SPR items (DIF statistics for three items were not calculated because for these items the percent correct in the focal group was zero; the MH D-DIF statistics cannot be calculated under this condition). The distribution for the fourth pretest administration shows the most

extreme negative DIF, with 21% of the items having MH D-DIF values of -1.5 and below. A frequency distribution of the DIF values for the four administrations combined is displayed in Figure 4. Overall, 80% of the DIF values for the pretest items are negative. The average MH D-DIF value for this set of SPR items is -.72 (SD = .88). The majority of MH D-DIF values fall in the interval between -1.0 and -0.5. However, of the 617 items, 101 have DIF values between -1.5 and -3.5. Thirty-eight items (6.1%) exhibit sizable DIF (exceed 1.5 in absolute value and are statistically different from 1.0 at the .05 significance level). None of the items in the sizable DIF category were selected for use in a final form.

Relationship Between Pretest and Final Form DIF Statistics

One-hundred pretest items were selected for inclusion in ten final forms, providing an opportunity to compare the DIF statistics across the two kinds of administrations. Note that the DIF statistics in the comparison are based on different kinds of samples. The items in the pretests were administered to test takers in high schools recruited for the field trials where individual scores were not reported. In contrast, the items in the final forms were administered to motivated test takers at national administrations. Also, the sample sizes for the African American group were considerably smaller in the pretest administrations than in the final form administrations. Finally, different matching criteria were used in the pretest and final form analyses.

Male/Female comparison.

Figure 5 displays the scatter plot of pretest MH D-DIF values and final form MH D-DIF values for 100 SPR items (a subset of the 620 items from the pretests). The correlation between MH D-DIF values from the two analyses is .75. The mean MH D-DIF shifted by .18 between the pretest administration

(mean = -.09) and the final form administration (mean = .09). The high correlation and the relatively small mean shift are an indication that the DIF values remained fairly consistent from the pretest administration to the final form administration.

White/African American Comparison.

Figure 6 displays the scatter plot of pretest MH D-DIF values versus final form MH D-DIF values for 100 SPR items (again, a subset of the 617 pretest items). The correlation between MH D-DIF values from the two analyses is .57, indicating a moderate positive relationship between DIF statistics from pretest and final form administrations. Many of the negative MH D-DIF values for the pretest items became less negative when the items were administered in a final form. The mean MH D-DIF shifted by .28 between the pretest administration (mean = -.59) and the final form administration (mean = -.31). Thus, DIF values for the final form items were less extreme than the DIF values for the same items administered to different test takers in field trial pretests.

Corrected Correlations.

Because the pretest statistics and the final form statistics are based on different sample sizes, the aforementioned correlations were corrected for unreliability, using formulas 14 and 15 in Zwick, Thayer, & Wingersky (1994)[1]. Corrected correlations were computed for 79 of the 100 items for which standard errors of the MH D-DIF statistic were readily available.

For the Male/Female comparison, the uncorrected and corrected correlations between the pretest and final form DIF statistics are .75 and .86, respectively. The modest difference between the uncorrected and corrected

[1]The correction procedure used in the Zwick et al (1994) paper was originally suggested by Charlie Lewis.

correlations is reasonable given the fairly large pretest and final form samples used to compute the MH D-DIF statistic for the Male/Female comparison.

For the White/African American comparison, the uncorrected and corrected correlations between the pretest and final form DIF statistics are .61 and 1.21 respectively. According to Zwick, Thayer, & Wingersky (1994), if reliability is underestimated, the corrected correlation can exceed unity; this appears to be the case for the White/African American comparison. Nevertheless, the corrected correlation suggests that there is good agreement between the White/African American DIF statistics obtained from the pretest administrations and the final form administrations.

Effect of the Matching Criterion on DIF Statistics (Selected Final Form Items)

The effect of the matching criterion was investigated to determine whether the presence of SPR items in the criterion alters DIF on these items. DIF statistics in four of the ten final forms were re-computed using a criterion that contained only the multiple choice items. Tables 5 and 6 display the item-level results and average DIF value by test form, for the analyses based on a criterion that included the SPR items (internal criterion) and excluded the SPR items (external criterion).

Male/Female comparison.

Table 5 shows that the analysis based on a criterion without SPR items resulted in DIF statistics that shifted very slightly in the positive direction. This finding is consistent with the finding that the SPR items exhibit a slight amount of positive DIF for females; a matching criterion based only on multiple choice items would tend to make the DIF on the SPR items slightly more extreme in the positive direction.

White/African American comparison.

Table 6 shows that the analysis based on a criterion without SPR items had a negligible effect on DIF statistics based on the total test.

Relationship Between DIF and Item Type (Final Form Items)

DIF statistics for the items that appeared in the ten final forms were broken down by item type.

Male/Female comparison.

Average DIF values for the three item types in SAT I are shown in Figure 7. Each data point represents the mean DIF value for items of one of the three types and the means are presented for each of the ten final forms. In general, the relationship between Male/Female DIF and item type is inconsistent. With the exception of two test forms (4 and 5), the average DIF values for the three item types are basically similar to each other. The average DIF value for the SPR items over the ten forms is .09, compared to -.02 for the five-choice multiple choice items and -.02 for the four-choice QC items.

White/African American comparison.

Figure 8 displays average DIF values for the White/African American comparison broken down by item type and test form. In contrast to the results for the Male/Female comparison, there is a consistent relationship between DIF and item type. Specifically, the average DIF value for the SPR items in each form is lower than the average DIF values for the other two item types. The average DIF value for the QC items is consistently higher than the average DIF values for the five-choice and SPR items. Average DIF values for the set of SPR items in each of the ten forms range between -.58 and -.14. The average DIF value for the SPR items over the ten forms is -.31,

compared to -.04 for the five-choice items and .12 for the four-choice QC items.

Relationship Between DIF and Item Difficulty (Final Form Items)

The relationship between DIF on the SPR items and item difficulty (based on the total group) was evaluated. The index of item difficulty is the expected item formula score (EIFS).[2] For SPR items, this index ranges between 0 and 1.0. For four-choice QC items, the expected item formula score ranges between -.33 and 1.00. For five-choice items, the expected item formula score ranges between -.25 and 1.00.

Male/Female comparison.

A scatter plot of DIF and item difficulty for the SPR items is shown in Figure 9. Although there is a slight tendency for easier items to exhibit more negative DIF, the magnitude of DIF is fairly consistent across item difficulty levels. The correlation between DIF and difficulty is .17. As a context for interpreting this scatter plot, data for the multiple choice item types are provided in the Appendix (Figures A.1 and A.2). The pattern of DIF and difficulty of SPR items is similar to the pattern for the multiple choice items.

White/African American comparison.

Figure 10 shows a scatter plot of DIF and item difficulty for the SPR items for the White/African American comparison. Similar to the findings

---

[2] Expected item formula score is computed as follows:

$$EIFS_i = \Sigma_s[((PR_{is}) - (1/(k_i-1)) * (PW_{is})) * (CBP_s)]$$

PR_{is} =percent right for each item at each scaled score level.
PW_{is} = percent wrong for each item at each scaled score level.
CBP_s = percent of examinees at each scaled score in the 1990 college bound senior cohort
k_i = the number of response options for each item.

for the Male/Female comparison, the magnitude of DIF is fairly consistent across item difficulty levels. The correlation between DIF and difficulty is -.19. A similar pattern is found for the multiple choice items (see Appendix, Figures A.3 and A.4), although negative DIF (with the exception of one item) is absent for very easy QC items (EIFS <.2).

Relationship Between DIF and Item Discrimination

The relationship between DIF on the SPR items and item discrimination was also evaluated. The index of item discrimination is the biserial correlation (rbis) between the item and the total test (in the total group).

Male/Female comparison.

A scatter plot of DIF and item discrimination for the SPR items is shown in Figure 11. Note that the discrimination values for the SPR items range between .40 and .90. The more discriminating items (rbis > .60) tend to exhibit more extreme DIF in both negative and positive directions. The less discriminating items (rbis < .60) tend to exhibit positive DIF only. The correlation between DIF and discrimination is -.02. The relationship between DIF and item discrimination for the multiple choice item types is shown in the Appendix. Although the relationship is not strong, the DIF values for the more discriminating (rbis > .50) five-choice items tend to be more variable than the DIF values for the less discriminating (rbis < .50) five-choice items (Figure A.5). However, for the 4-choice QC items, the magnitude of DIF is consistent across levels of item discrimination (Figure A.6).

Figure 12 provides another way to look at the relationship between DIF, item discrimination, and item type. Each of the three panels in this figure is a scatter plot of average DIF and average item discrimination, where

the average is the mean DIF value for the ten SPR items in each of the ten final forms. In the context of the multiple choice items, it is clear from this figure that the SPR items are much more discriminating. Compared to the multiple-choice items, these more discriminating SPR items tend to exhibit DIF that is slightly positive for female test takers.

White/African American comparison.

Figure 13 shows a scatter plot of DIF and item discrimination for the SPR items for the White/African American comparison. Within the SPR item type, the magnitude of DIF is fairly consistent across levels of item discrimination. The correlation between DIF and discrimination is .10. However, the less discriminating items (rbis < .60) tend to exhibit slightly less DIF, particularly, less positive DIF. A similar pattern is found for the multiple choice items (see Appendix, Figures A.7 and A.8), although Figure A.7 indicates that there is a slight tendency for the more discriminating five-choice items to exhibit more DIF than the less discriminating five-choice items. Figure 14 shows scatter plots, by item type, of average DIF and average item discrimination. An item type effect is evident in this figure -- the more discriminating SPR items tend to exhibit more negative DIF for African American test takers, relative to the less discriminating multiple choice items.

Discussion and Summary

The present research summarized DIF results for SPR items, a non-multiple choice mathematical item type in the new SAT I. This new item type was introduced in the SAT I for the first time in March 1994. DIF data from field trial pretests (620 SPR items) and final forms (100 SPR items) were examined.

For the Male/Female comparison, the SPR item type exhibits negligible DIF. The Male/Female DIF results for the SPR items resemble DIF results for the multiple choice items in the SAT.

For the African American/White comparison, SPR items exhibit more DIF than the multiple choice items. (Moderate or sizable DIF on multiple choice items in the old or new SAT-M is rare.) In particular, when African American test takers are compared to matched groups of White test takers, they tend to perform less well than White test takers on the SPR items. Although the level of African American/White DIF on SPR items is elevated, relative to the multiple choice items, none of the SPR items in any of the final forms were flagged for exhibiting extreme DIF.

The finding of elevated DIF on SPR items is consistent with results reported by Schmitt and Crone (1991) based on prototype tests. However, in the Schmitt and Crone study, DIF on the SPR item type was observed for Females as well as for African Americans. It is difficult to compare results between the present research and the Schmitt and Crone research because the research used different matching criteria and the samples may have been quite different.

The consistent relationship between item type and DIF in the African American/White comparison signals some multidimensionality in the test. Thus, we may conclude that the item types measure mathematical ability somewhat differently for African American test takers and White test takers. A question to answer is whether the multidimensionality is related to something irrelevant to the construct of interest, or whether the difference in dimensional structure reflects a valid subcomponent of what the SAT I intends to measure.

DIF statistics were fairly stable between the pretest and final form analyses, particularly for the Male/Female comparison, where the correlation between pretest DIF statistics and final form DIF statistics was .75 (and a systematic difference of .18). For the Male/Female comparison, DIF on the SPR items tended to be negligible in both the pretest and final form analyses.

For the African American/White comparison, the correlation between pretest DIF statistics and final form DIF statistics was lower (.57). The pretest DIF statistics for the African American/White comparison shifted in the positive direction by .28 but tended to remain negative in the final form analyses (i.e., the DIF values became less extreme).

The shift in DIF statistics between pretest field trials and final form administrations may be explained by a variety of factors. One explanation may be related to the background of the test takers vis-a-vis preparation for the new item type and familiarity with the directions and procedures for using the answer sheet for gridding responses to the SPR items. Preparation materials were distributed to the high schools prior to the field trial administrations. However, it is likely that the use of these preparation materials was inconsistent from school to school. This possibility is noteworthy because some of the schools that volunteered to participate in the pretest field trials had very high proportions of African American test takers; other schools had very low proportions of African American test takers. Preparation materials were also distributed to the high schools prior to the final form administration. Again, it is possible that the use of these preparation materials was inconsistent from school to school. It is also possible that pretest test takers and final form test takers used the preparation material to varying degrees. One might speculate that there was more motivation to read the preparation materials for the final form

administration (where scores were to be reported to students and institutions) compared to the pretest administration where performance on the test did not result in a score report for a student. The finding of less extreme DIF on SPR items in the final form analyses is consistent with this hypothesis.

An important difference between the DIF analyses on items in pretests and the final forms was sample size. The sample sizes for the pretest analyses (particularly in the African American/White comparison) were considerably smaller than the sample sizes for the final form analyses. Consequently, the DIF statistics are expected to be less stable in the smaller samples and this would reduce the correlation between the pretest statistics and the final form statistics. When corrected for unreliablity, the correlations indicated good agreement between pretest and final form DIF statistics.

In addition to differences in sample sizes, different matching criteria were used to compute the DIF statistics for the pretest analyses (external, based on four-choice QC items) and the final forms analyses (internal, based on four-choice QC items, five-choice items, and SPR items). Whereas, the external criterion used in the pretest analyses was a 25-item test, the internal criterion in the final form analyses was a 60-item test. Uttaro and Millsap (1994) have shown that the number of strata in the matching criterion may have an effect on consistency of DIF results across analyses.

The content of the matching criterion has also been shown to affect DIF results (e.g., Ryan, 1991; McPeek and Wild, 1986). However, the content of the matching criterion was found to have little impact on DIF results in an earlier study of SPR items (Schmitt & Crone, 1991). A similar result was found in the present research. Nevertheless, the inconsistent DIF statistics for the African American/White comparison between pretest and final form may be partially related to the finding that QC items tend to have positive DIF

for African American test takers, and the matching criterion used in the pretest analyses was composed entirely of QC items.

A clear finding from this research is that the SPR items are more discriminating than the multiple choice items and that the SPR items exhibit more negative DIF for African Americans than the multiple choice items. This finding is consistent with previous simulation studies of factors affecting DIF measurement (Donoghue, Holland & Thayer, 1993). The portion of the Donoghue et al. study that looked at factors affecting MH D-DIF showed that item discrimination levels interacted most strongly with two factors -- the amount of DIF in the studied item and the difficulty of the studied item. They show that in a Rasch model with constant slope a, MH D-DIF provides an estimate of $4a\,(b_R - b_F)$, where b_R and b_F are the reference and focal group difficulties, respectively. Thus, there is a theoretical basis for the relationship between DIF and discrimination (given a fixed difference in difficulty for reference and focal groups), such that, for a given difference in difficulties, a has a multiplicative effect on the size of the DIF statistic. The results in the present research are also consistent with findings of Zwick, Thayer, and Wingersky (1994), who show that the relationship between DIF and discrimination is consistent with an IRT formulation of DIF.

Given the theoretical basis for the relationship between item discrimination and DIF, and the systematic finding in this research, the relationship between item discrimination and DIF on SPR items seems to warrant further exploration. While multiple choice items provide an opportunity for low ability test takers to guess the correct answer, the opportunity to guess the correct answer on an SPR items is effectively eliminated. Thus, it is not surprising that SPR items are more discriminating than multiple choice items.

While item format may explain the higher item discrimination for the SPR items, it is also relevant to point out Masters' (1988) concerns that high discrimination for items sometimes occurs for the wrong reason. In addition, Yen (1980) observed that items at the end of a section can have higher discrimination as a result of a second dimension (which she referred to as carelessness and Masters referred to as frustration) that is negatively correlated with ability. In the context of the present research, perhaps SPR items give lower scoring test takers an additional disadvantage because the item type appears at the end of the test section. Note that African American test takers tend to score lower than White test takers on SAT-M.

All of the research on SPR items involves administration of this item type at the end of a test section. Using data from SAT verbal forms, Dorans, Schmitt, and Curley (1988) carried out research to determine whether a location effect on DIF results could be removed statistically. Dorans et al concluded that the procedure of excluding test takers who do not reach an item from the calculation of the DIF statistic for that item partially compensates for the effects of item location on the DIF estimate. However, the research involved multiple choice items, where a contiguous string of omit responses to items at the end of a section may reasonably be interpreted as not-reached responses (assuming that test takers pace through the items in each section in the order that the items are presented and that the omit response did not represent examinees who attempted to answer the question but could not). SPR items reflect a situation that differs from the multiple choice situation, because test takers who do not know the answer to SPR items at the end of the test will tend to leave the answer sheet blank rather than attempt an answer from among the choices. Research is needed to determine whether the DIF statistic for the SPR item type is appropriately

adjusted for item location effects by excluding not-reached responses from the computation of the statistic (this was the procedure used in the present research). Research on manipulating the item position of the SPRs may prove useful to observe the effects on DIF and item discrimination.

A major finding from this research is that SPR items tend to exhibit negative DIF for African American test takers, although the DIF statistics for individual items are not extreme in the final forms. Nevertheless, each edition of the test contains ten SPR items, and DIF for an item type that consistently exhibits negative DIF for a particular focal group needs to be monitored. In addition, the relationship between higher item discrimination and DIF needs to be understood from the standpoint of what the test is intended to measure. Linn (1993) demonstrated a strong relationship between DIF and item discrimination. To use Linn's words: "Items that are related most strongly to total test scores are also most likely to be flagged for having large negative DIF values" (p.363). It may be the case that the SPR items exhibit more negative DIF than the multiple choice items because the SPR items provide a clearer picture of group differences than the test as a whole.

The research described in this paper utilized the Mantel-Haenszel methodology to assess differential item functioning. However, Roussos and Stout (1993) showed that the Mantel-Haenszel method can have false-positive flagging problems when the studied item is a lot more discriminating than the items in the matching criterion. Uttaro and MIlsap (1994) also show that item parameter values can influence MH D-DIF results.

The effect of using alternative DIF assessment procedures is worth exploring. Research by Mazzeo and Chang (1994) shows that the SIBTEST procedure, proposed by Shealy and Stout (1993), is more robust than the

Mantel-Haenszel procedure when the discrimination parameter of the studied item differs considerably from the average discrimination of the items in the matching criterion. For their simulation study, ability distributions for the focal group and reference group differed by one standard deviation, which is similar in magnitude to the difference between score distributions for White and African American test takers in the present research. Given the high reliability of the matching criterion used in the present research, large differences between results based on the Mantel-Haenszel procedure and the SIBTEST procedure are not expected. However, a worthwhile follow-up study to the present research is to use SIBTEST to analyze DIF on the SPR items.

References

Braswell, J. (1991, April). Rationales for current and proposed SAT mathematics item types. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

College Board (1994). 1994 profile of SAT and Achievement test takers.

Donoghue, J.R., Holland, P.W., & Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 137 - 166). Hillsdale, NJ: Erlbaum.

Dorans, N.J., Schmitt, A.P., & Curley, W.E. (1988, April). Differential speededness: Some items have DIF because of where they are, not of what they are. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Fremer, J.J. (1993, April). What should determine the content of the SAT? Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Holland, P.W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In. H. Wainer & H.I. Braun (Eds.), Test validity, (pp. 129-145). Hillsdale, NJ: Erlbaum.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 349 - 364). Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Masters, G.N. (1988). Item discrimination: When more is worse. Journal of Educational Measurement, 25, 15-29.

Mazzeo, J. & Chang, H. (1994, April). Detecting DIF in polytomously-scored items: Progress in adapting Shealy and Stout's SIBTEST. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

McPeek, W.M. & Wild, C.L. (1986, April). Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Roussos, L. & Stout, W. (1993, April). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Ryan, K.E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. Journal of Educational Measurement, 28, 325-337.

Schmitt, A. & Crone, C. (1991). Alternative mathematical aptitude item types: DIF issues. (RR-91-42). Princeton, NJ: Educational Testing Service.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58, 159-194.

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18, 15-25.

Yen, W.M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

Zwick, R., Thayer, D. & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. Applied Psychological Measurement, 18, 121-140.

Table 1
Configuration of Field Trial Pretests and Final Forms

Administration	Section	Item Position	Item Type
Field Trial Pretest	1	1-25	4-choice QC
		26-35	SPR
Final Form	1	1-25	5-choice
	2	1-15	4-choice QC
		16-25	SPR
	3	1-10	5-choice

Table 2

Ranges of Sample Sizes (Per Pretest) for DIF Analyses of SPR Items in Field Trial Pretest Administrations

Pretest Admin.		Males	Females	Whites	African Americans	No. of Pretests	No. of Items
First	Min.	998	1,173	1,653	288	10	100
	Max.	1,247	1,378	1,905	362		
Second	Min.	470	510	762	133	12	120
	Max.	554	639	802	247		
Third	Min.	589	687	947	129	20	200
	Max.	716	845	1,154	190		
Fourth	Min.	1,583	1,768	2,528	417	20	200
	Max.	1,871	2,163	3,054	509		

Note: Samples for DIF analyses were restricted to college-bound juniors and seniors who identified themselves as having English as one of their first languages.

Table 3

Sample Sizes for DIF Analyses of SPR Items in Final Forms

Test Form	Males	Females	Whites	African Americans
1	67,878	80,037	112,995	15,996
2	61,996	75,437	109,713	9,864
3	30,020	37,021	53,936	4,836
4	30,008	36,888	53,579	4,734
5	33,693	40,428	56,233	7,035
6	33,382	40,665	56,171	7,020
7	33,556	40,608	56,349	6,937
8	72,856	104,272	128,234	16,170
9	124,381	165,062	222,098	29,648
10	63,697	74,872	94,418	20,986

Note: Samples for DIF analyses were restricted to college-bound juniors and seniors who identified themselves as having English as one of their first languages.

Table 4
Reliability of Matching Criteria

Matching Variable	Type	Reliability
25 Quantitative Comparison items	External	.81 - .85
60-item total test	Internal	.92 - .94
50 multiple-choice items	External	.90 - .93

Table 5

Results of DIF Analyses Using Internal and External Matching Criteria for the Male/Female Comparison

	Form 2			Form 5			Form 6			Form 9		
SPR	Int.	Ext.	Diff.	Int.	Ext.	Diff.	Int.	Ext.	Diff.	Int.	Ext.	Diff.
1	1.40	1.37	0.03	0.24	0.28	-0.04	0.19	.21	-0.02	0.09	0.03	0.06
2	0.05	-0.01	-0.04	0.68	0.71	-0.03	0.30	0.31	-0.01	-0.47	-0.41	-0.06
3	-0.22	-0.17	-0.05	0.24	0.27	-0.03	0.98	0.98	0.00	0.50	0.47	0.03
4	0.01	0.05	-0.04	-0.11	-0.02	-0.09	-0.14	-0.10	-0.04	0.48	0.49	-0.01
5	0.56	0.60	-0.04	0.40	0.45	-0.05	-0.41	-0.35	-0.06	-0.04	-0.04	0.00
6	0.66	0.70	-0.04	0.81	0.88	-0.07	0.85	0.88	-0.03	0.43	0.47	-0.04
7	- 0.13	-0.07	-0.06	-0.27	-0.18	-0.09	-0.68	-0.61	-0.07	-0.12	-0.09	-0.03
8	-1.03	-0.93	-0.10	-0.22	-0.13	-0.09	-0.53	-0.45	-0.08	0.02	0.09	-0.07
9	0.08	0.13	-0.05	0.55	0.62	-0.07	0.60	0.64	-0.04	0.21	0.20	0.01
10	0.11	0.16	-0.05	0.01	0.08	-0.07	-0.71	-0.65	-0.06	0.36	0.33	0.03
Mean	0.14	0.18	-0.04	0.23	0.30	-0.07	0.05	0.09	-0.04	0.15	0.15	0.00

Note: Int. = Internal 60-item criterion; Ext. = External 50-item criterion (no SPR items); the studied item was included in the criterion; Diff.=Difference (internal minus external) between MH values.

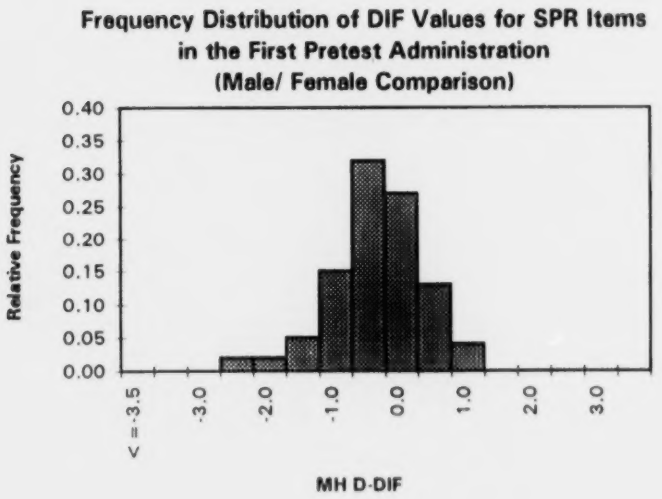
Table 6

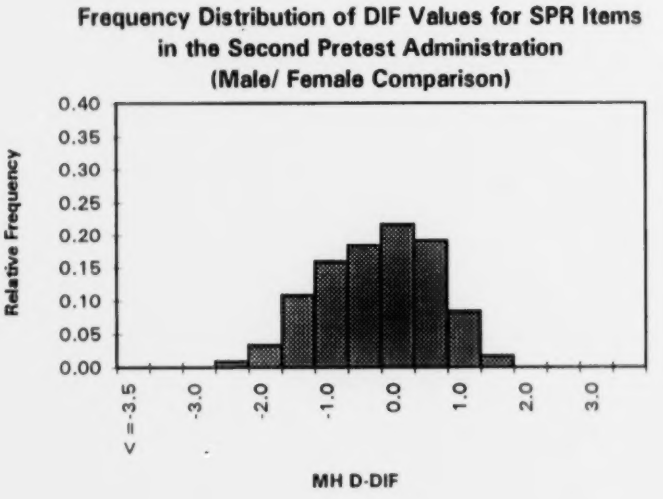
Results of DIF Analyses Using Internal and External Matching Criteria for the White/African American Comparison

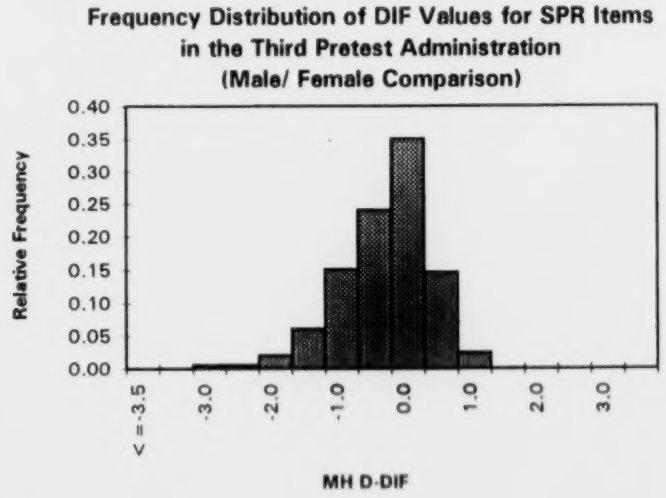
	Form 2			Form 5			Form 6			Form 9		
SPR	Int.	Ext.	Diff.	Int.	Ext.	Diff.	Int.	Ext.	Diff.	Int.	Ext.	Diff.
1	0.10	0.04	0.06	0.34	0.41	-0.07	0.32	0.33	-0.01	-0.78	-0.84	0.06
2	-0.74	-0.71	-0.03	0.11	0.12	-0.01	-0.48	-0.43	-0.05	-0.60	-0.61	0.01
3	-0.81	-0.77	-0.04	-0.65	-0.64	-0.01	0.47	0.47	0.00	-0.16	-0.24	0.08
4	-0.50	-0.47	-0.03	-0.14	-0.12	-0.02	-1.32	-1.26	-0.06	-0.02	0.02	-0.04
5	0.54	0.52	0.02	-0.30	-0.24	-0.06	0.02	0.04	-0.02	-0.04	-0.06	0.02
6	-0.23	-0.22	-0.01	0.48	0.49	-0.01	0.76	0.78	-0.02	0.10	0.01	0.09
7	-0.34	-0.34	0.00	0.01	0.05	-0.04	-0.08	-0.07	-0.01	-0.42	-0.40	-0.02
8	-0.62	-0.60	-0.02	-0.14	-0.12	-0.02	-0.86	-0.82	-0.04	-0.97	-1.07	0.10
9	-0.26	-0.25	-0.01	-0.16	-0.13	-0.03	0.17	0.18	-0.01	0.08	0.13	-0.05
10	-0.75	-0.74	-0.01	-1.13	-1.10	-0.03	-0.40	-0.42	0.02	0.39	0.37	0.02
Mean	-0.36	-0.35	-0.01	-0.16	-0.13	-0.03	-0.14	-0.12	-0.02	-0.24	-0.27	0.03

Note: Int. = Internal 60-item criterion; Ext. = External 50-item criterion (no SPR items); the studied item was included in the criterion; Diff.=Difference (internal minus external) between MH values.

Figure 1


Frequency Distribution of DIF Values for SPR Items
in the First Pretest Administration
(Male/ Female Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< = -3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF


Frequency Distribution of DIF Values for SPR Items
in the Second Pretest Administration
(Male/ Female Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< = -3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF


Frequency Distribution of DIF Values for SPR Items
in the Third Pretest Administration
(Male/ Female Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< = -3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF

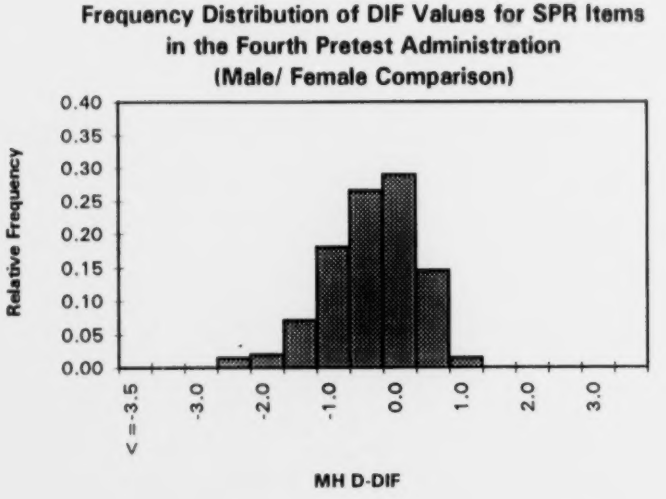

Frequency Distribution of DIF Values for SPR Items
in the Fourth Pretest Administration
(Male/ Female Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< = -3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF

Figure 2

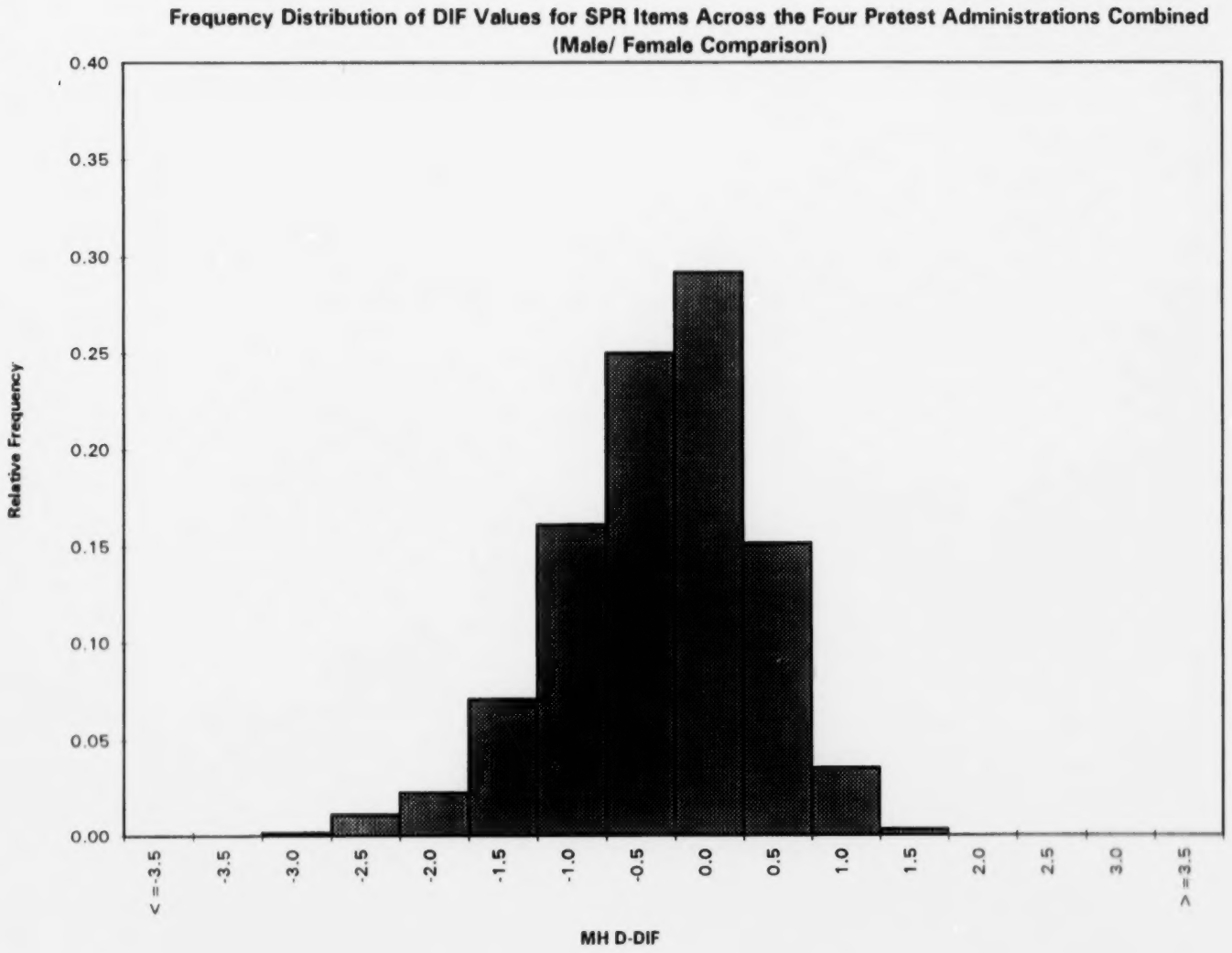
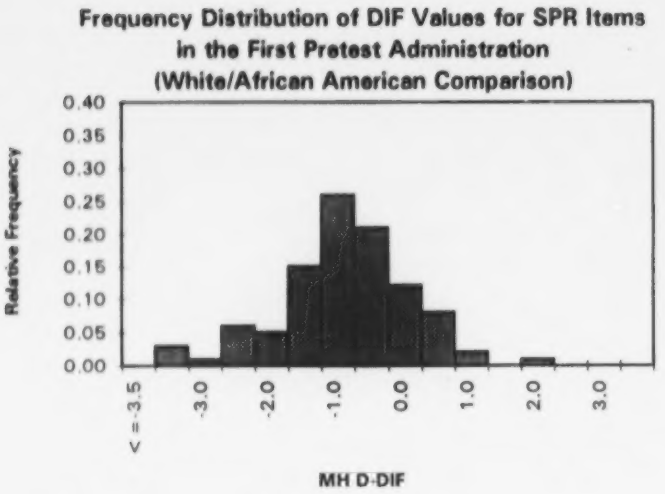
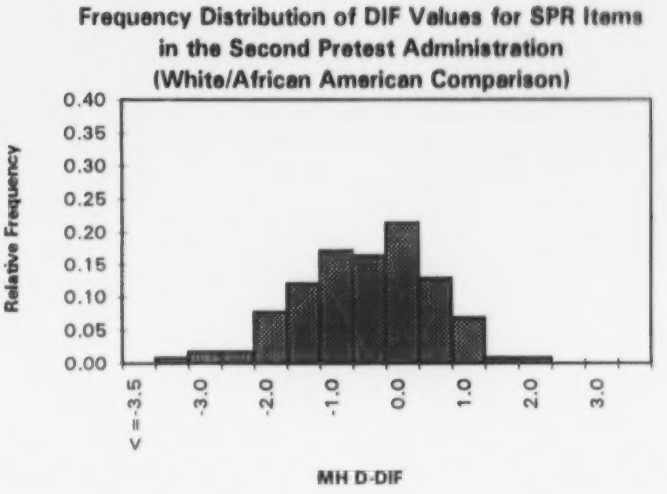
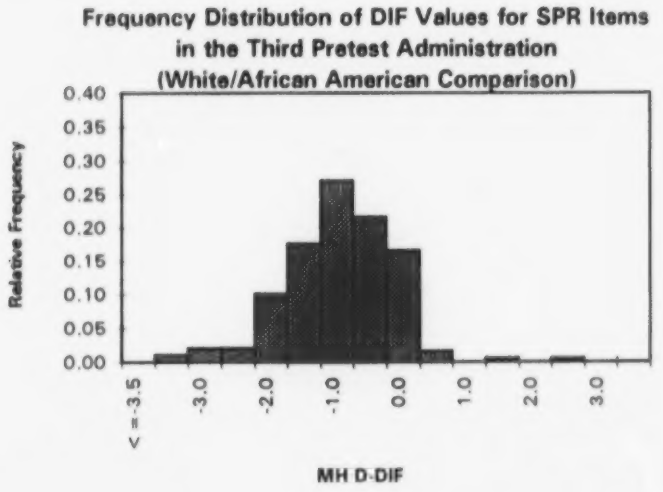

Frequency Distribution of DIF Values for SPR Items Across the Four Pretest Administrations Combined
(Male/ Female Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< =-3.5
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
> =3.5
MH D-DIF

Figure 3


Frequency Distribution of DIF Values for SPR Items
in the First Pretest Administration
(White/African American Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< =-3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF


Frequency Distribution of DIF Values for SPR Items
in the Second Pretest Administration
(White/African American Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< =-3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF


Frequency Distribution of DIF Values for SPR Items
in the Third Pretest Administration
(White/African American Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< =-3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF

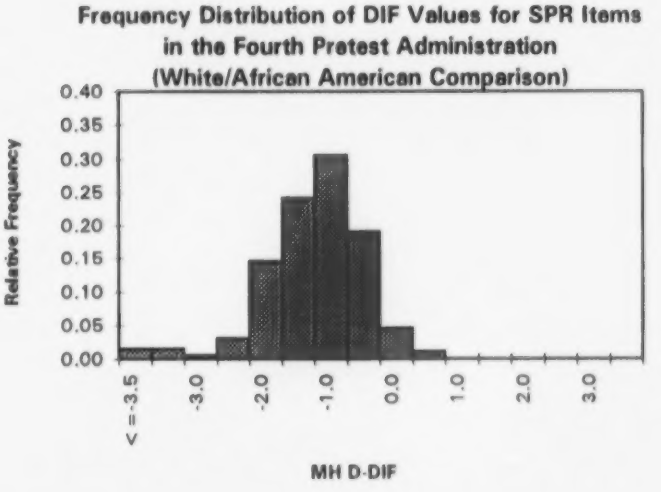

Frequency Distribution of DIF Values for SPR Items
in the Fourth Pretest Administration
(White/African American Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< =-3.5
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
MH D-DIF

Figure 4

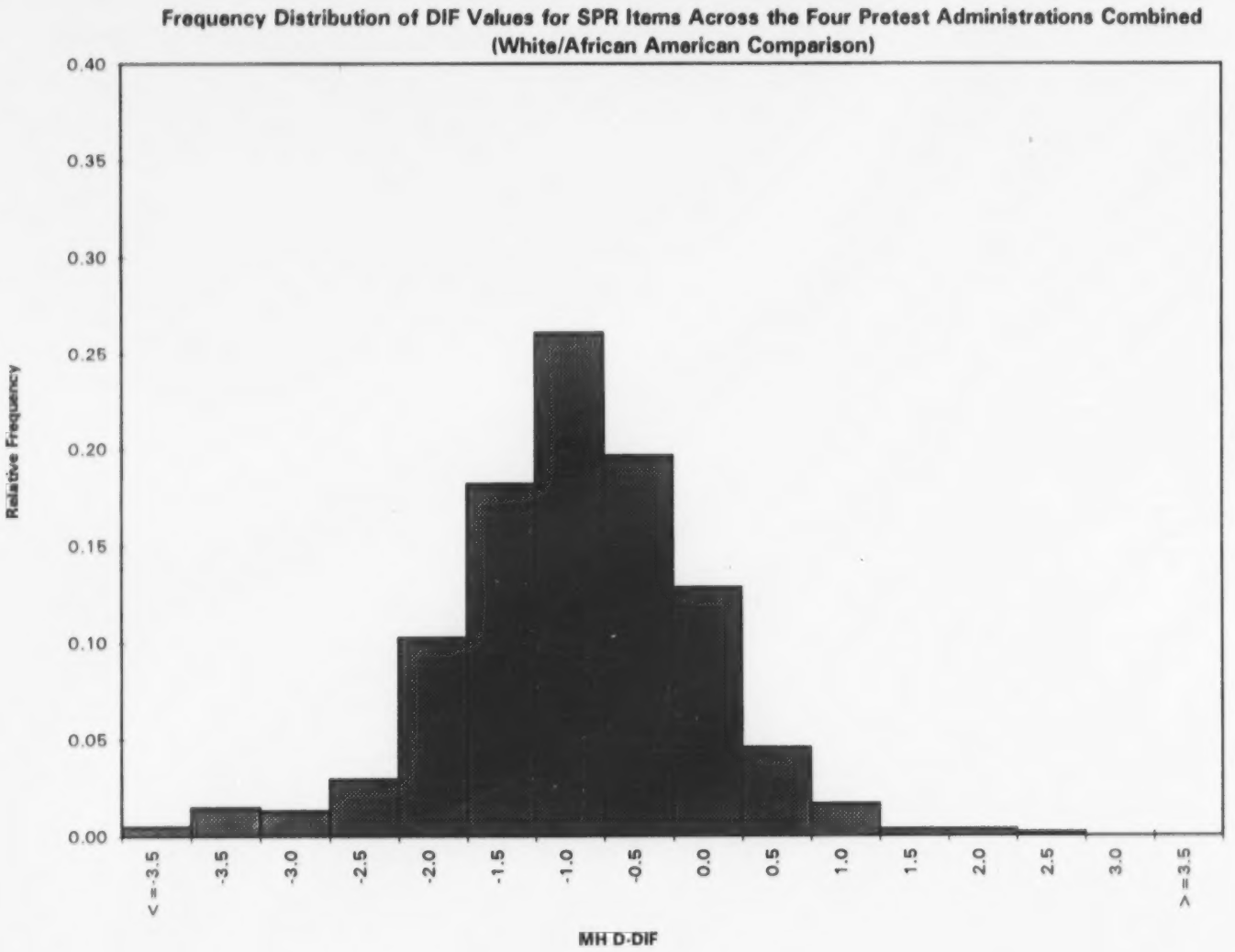

Frequency Distribution of DIF Values for SPR Items Across the Four Pretest Administrations Combined
(White/African American Comparison)
Relative Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
< = -3.5
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
> = 3.5
MH D-DIF

Figure 5

Scatter Plot of Pretest DIF and Final Form DIF
(Male/Female Comparison)

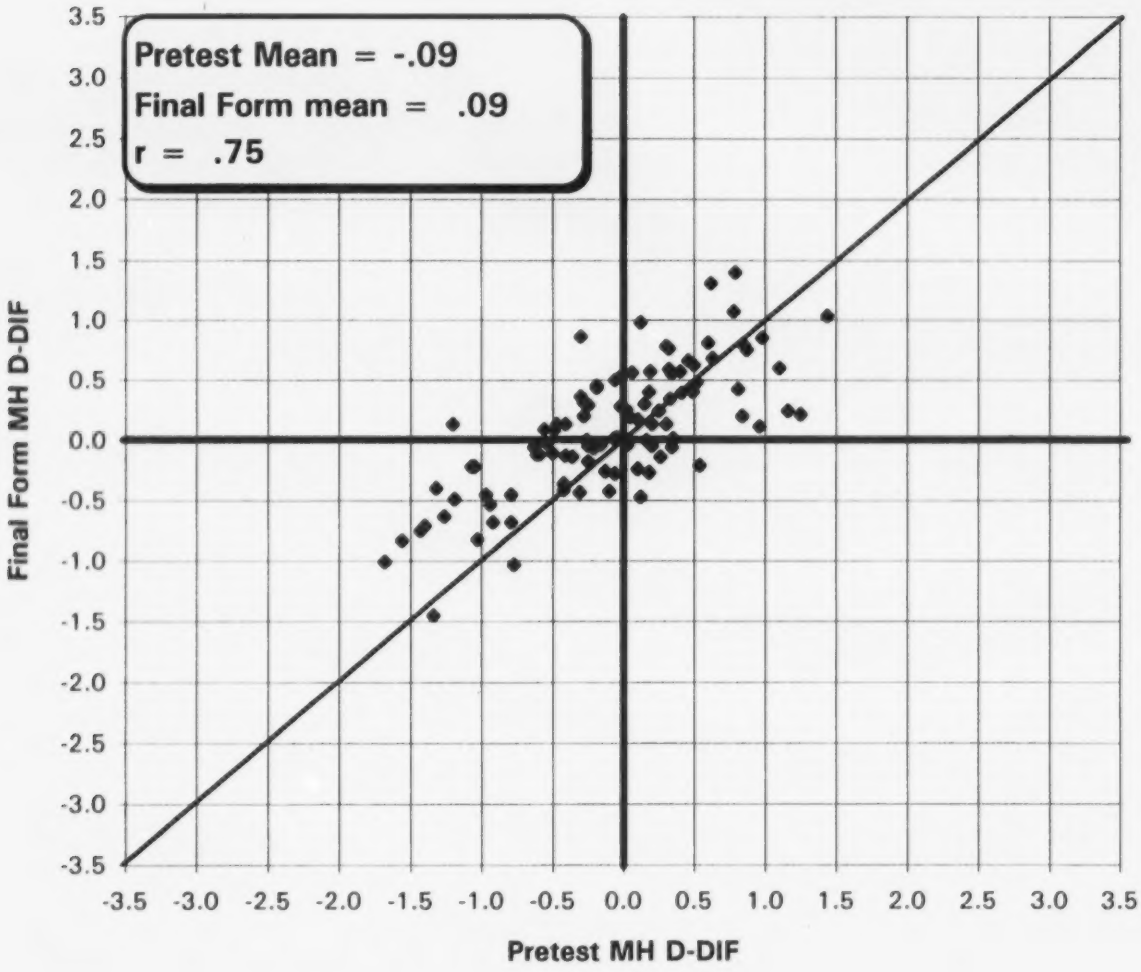


Figure 6

Scatter Plot of Pretest DIF and Final Form DIF
(White/African American Comparison)

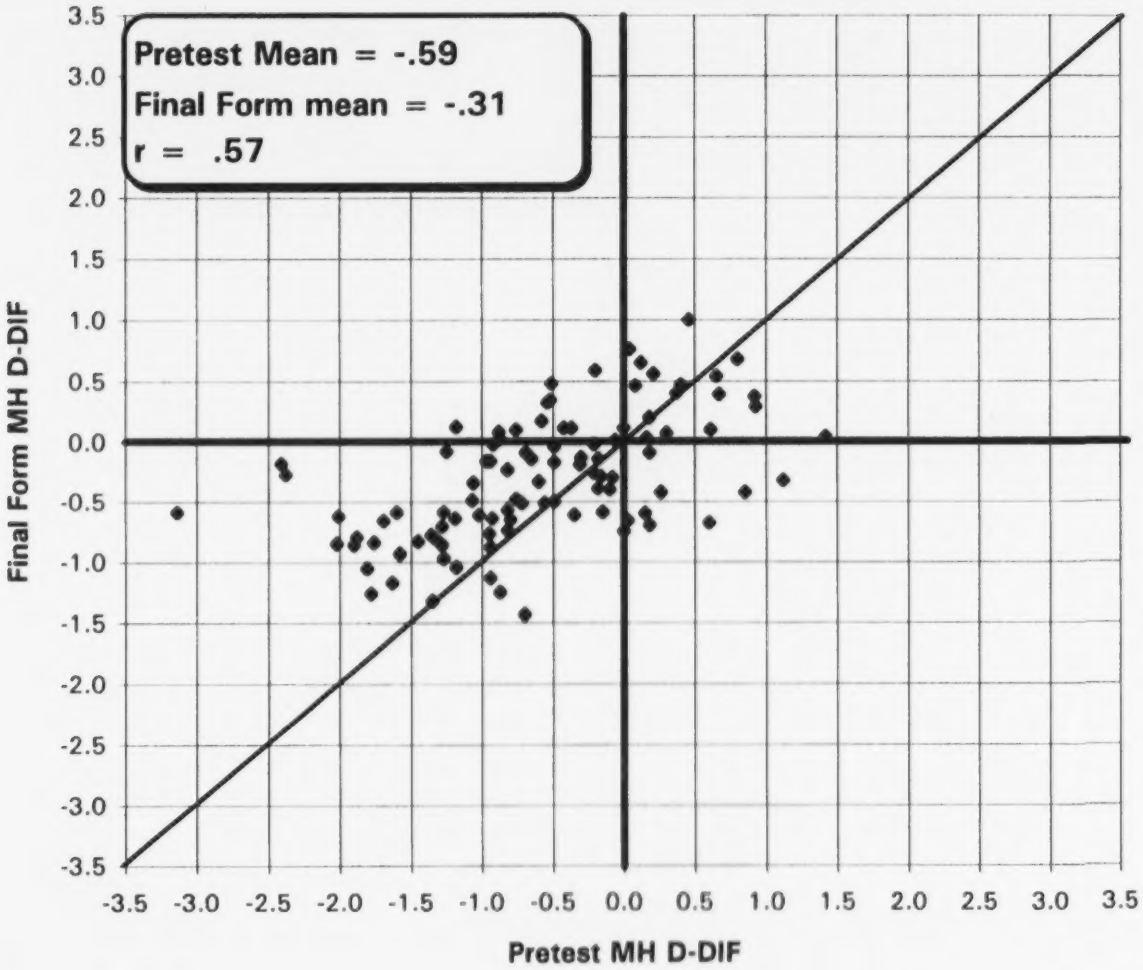


Figure 7

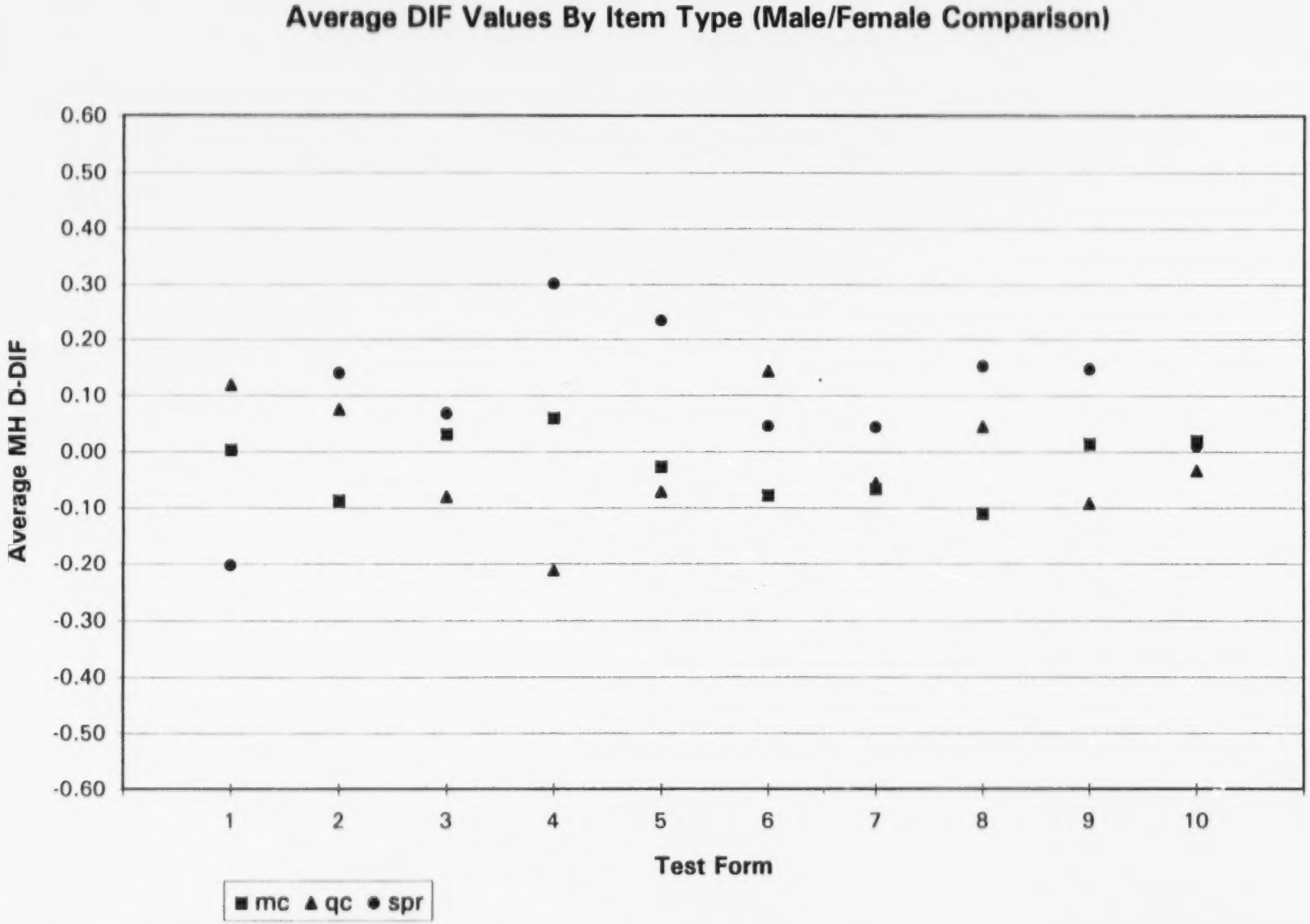

Average DIF Values By Item Type (Male/Female Comparison)
Average MH D-DIF
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
-0.40
-0.50
-0.60
1
2
3
4
5
6
7
8
9
10
Test Form
mc
qc
spr

Figure 8

Average DIF Values By Item Type (White/African American Comparison)

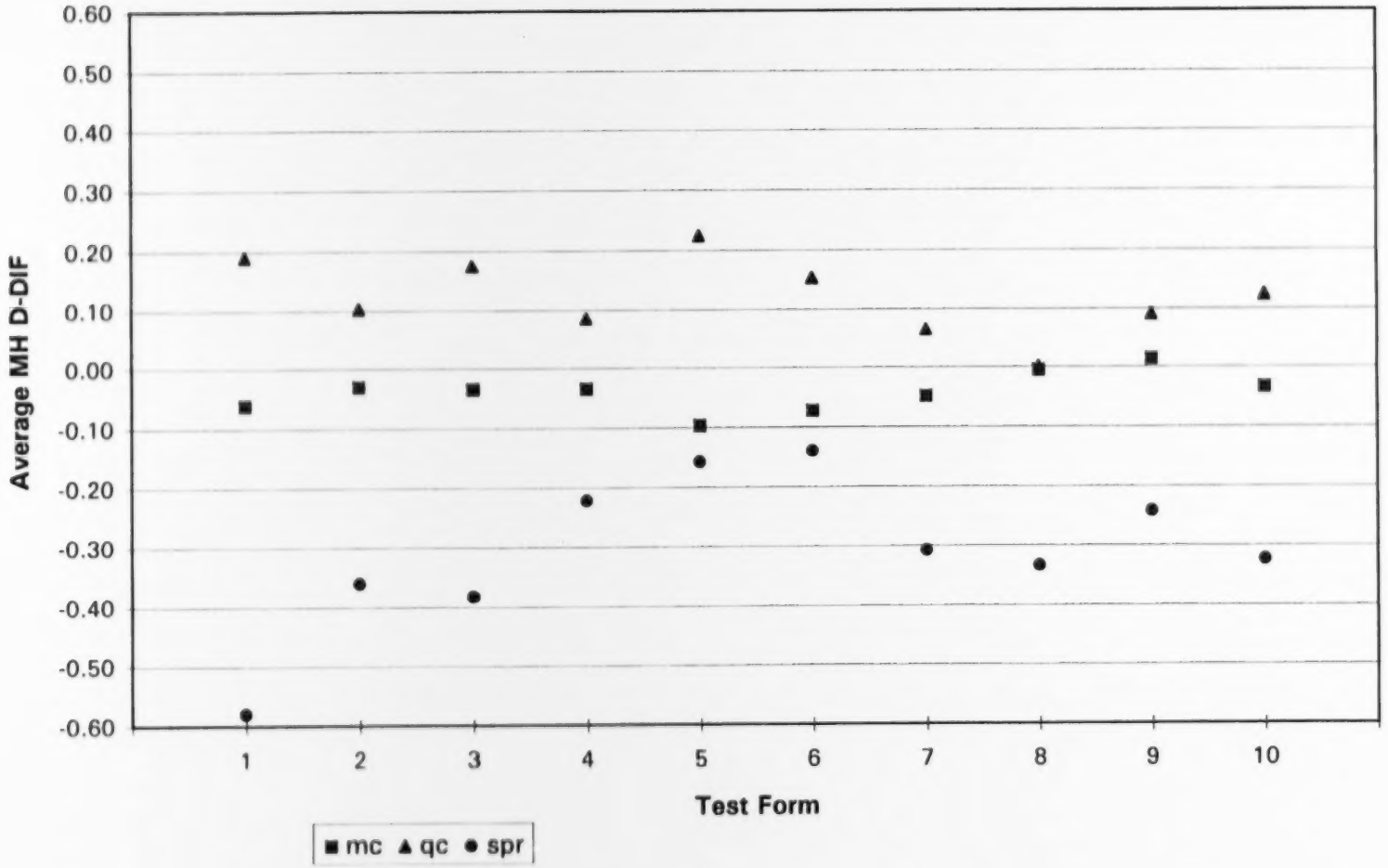


Figure 9

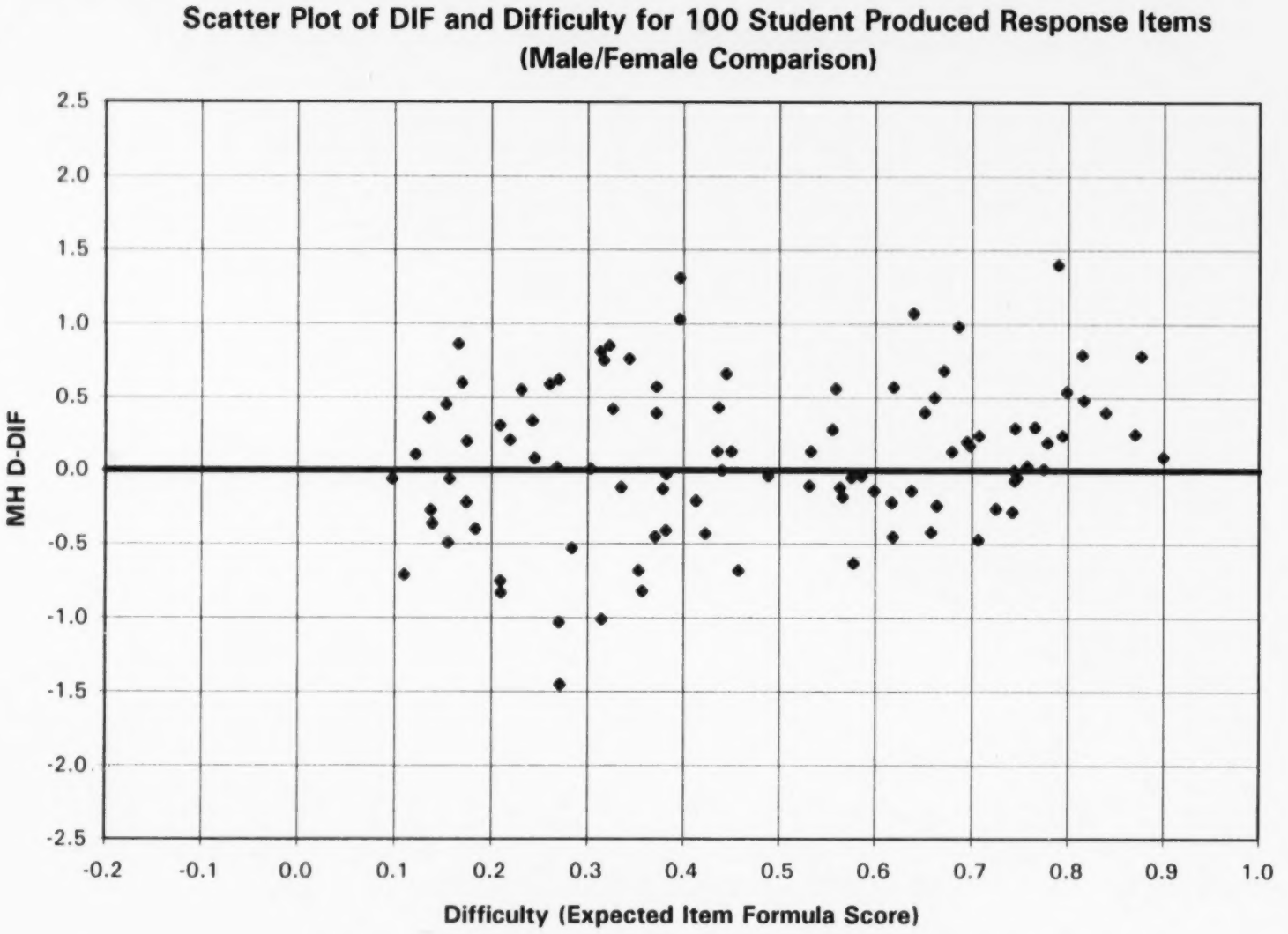

Scatter Plot of DIF and Difficulty for 100 Student Produced Response Items
(Male/Female Comparison)
MH D-DIF
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Difficulty (Expected Item Formula Score)

Figure 10

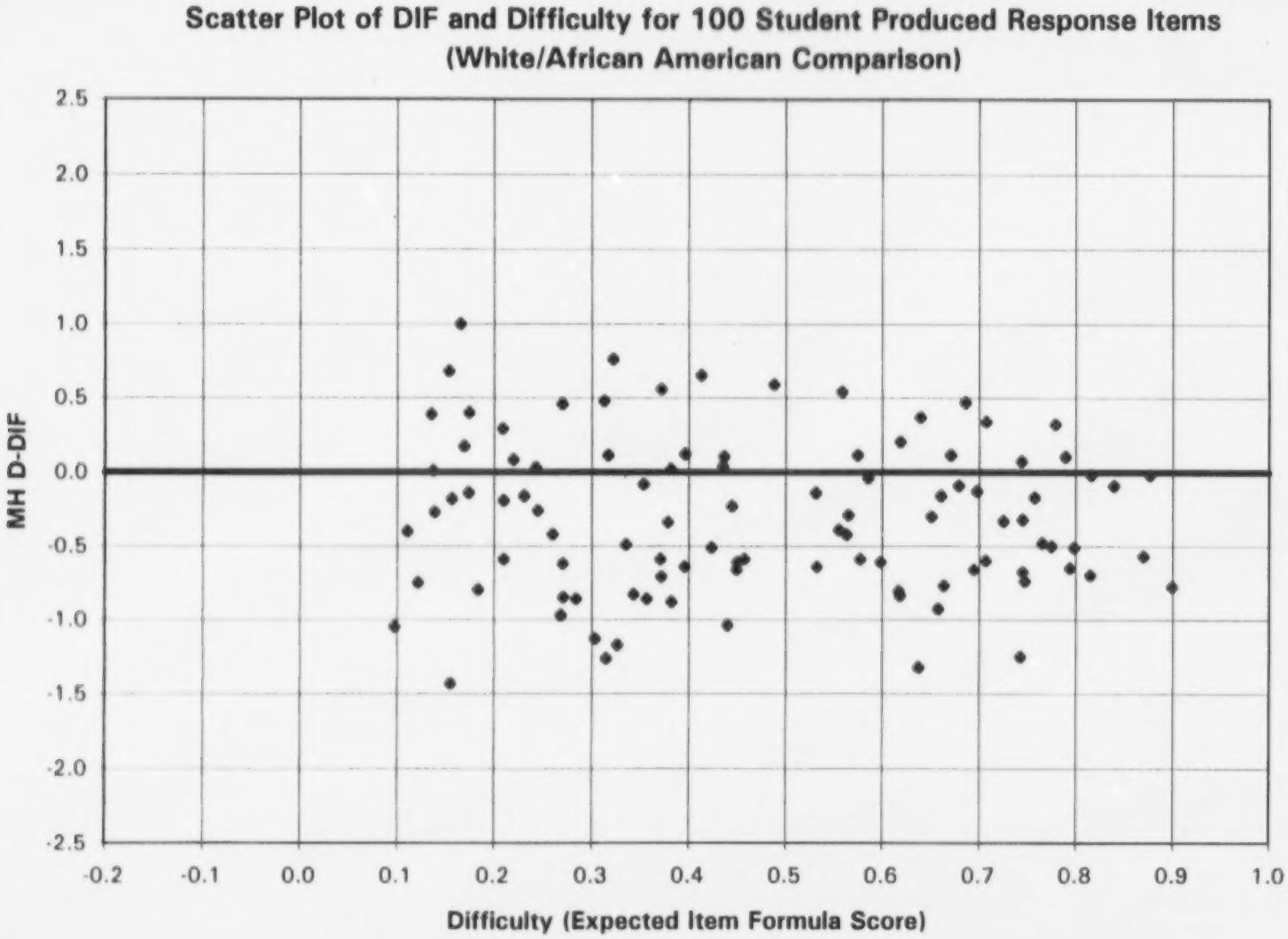

Scatter Plot of DIF and Difficulty for 100 Student Produced Response Items
(White/African American Comparison)
MH D-DIF
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Difficulty (Expected Item Formula Score)

Figure 11

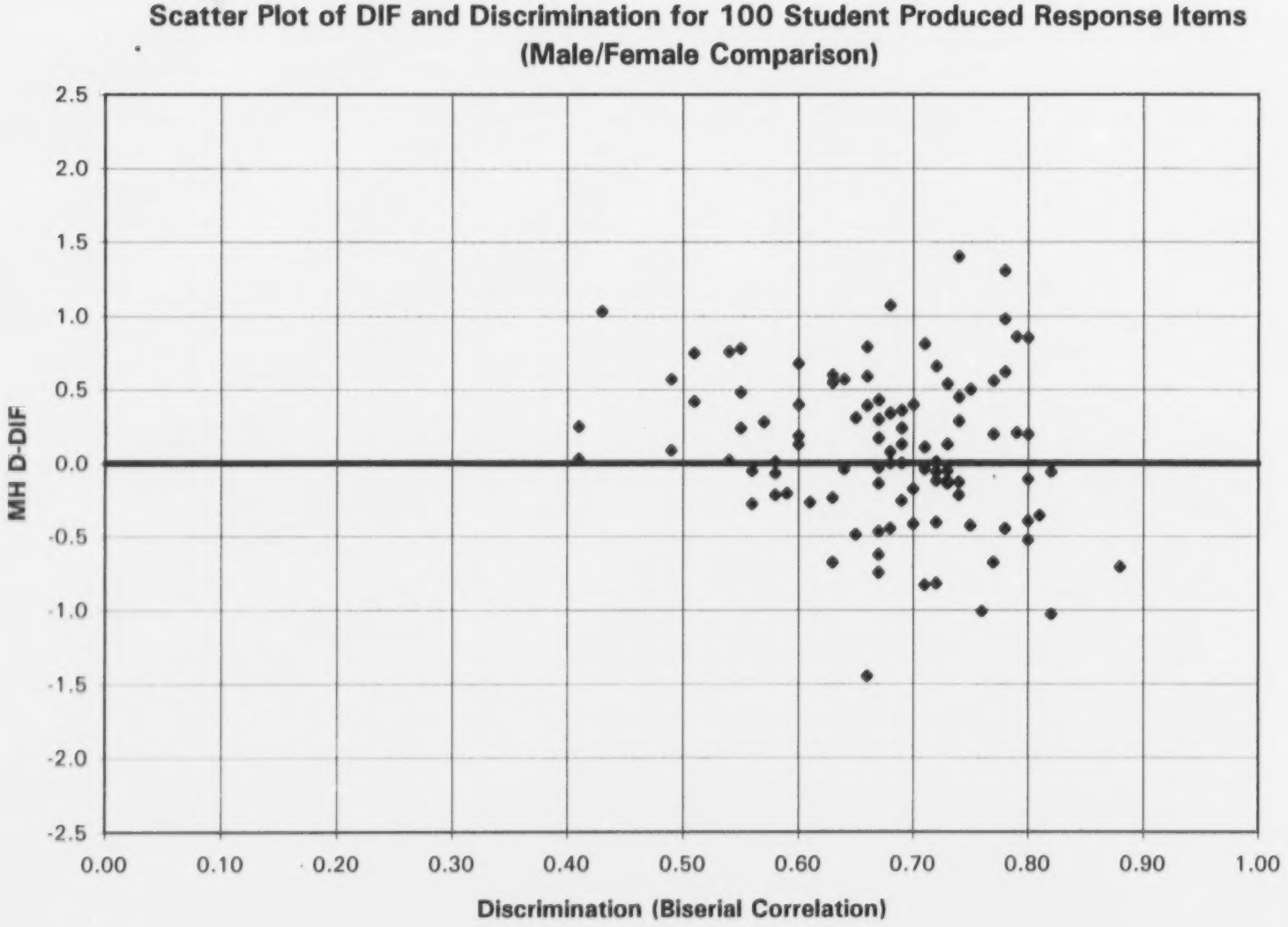
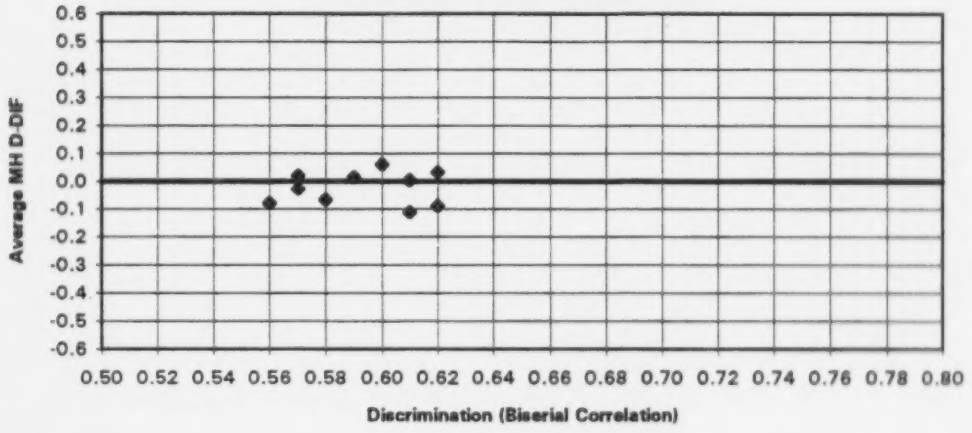
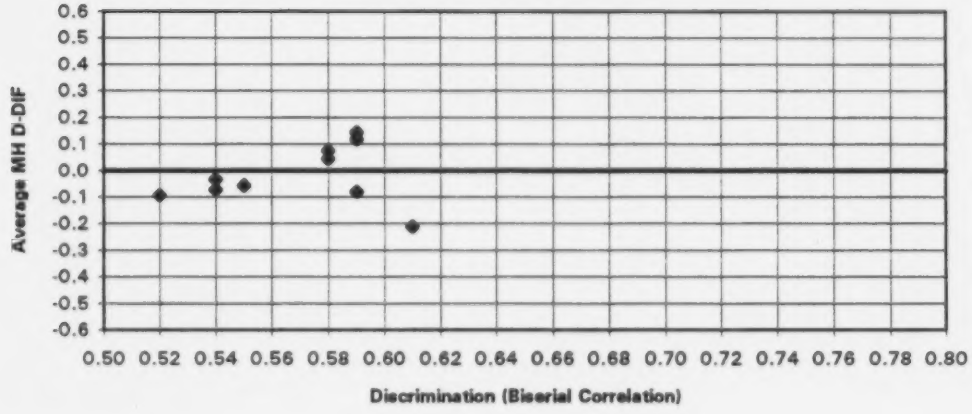

Scatter Plot of DIF and Discrimination for 100 Student Produced Response Items
(Male/Female Comparison)
MH D-DIF
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Discrimination (Biserial Correlation)

Figure 12

Scatter Plot of Average DIF and Discrimination Values by Form for 5-Choice Items (Male/Female Comparison)



Scatter Plot of Average DIF and Discrimination Values by Form for Quantitative Comparison Items (Male/Female Comparison)



Scatter Plot of Average DIF and Discrimination Values by Form for Student Produced Response Items (Male/Female Comparison)

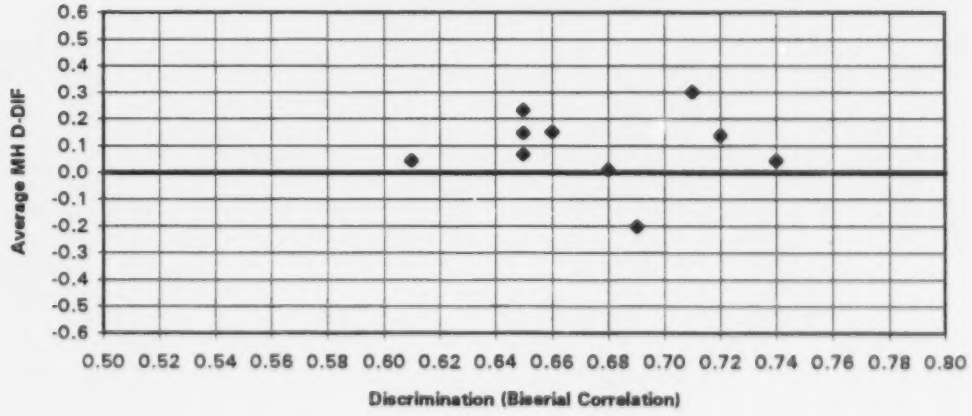


Figure 13

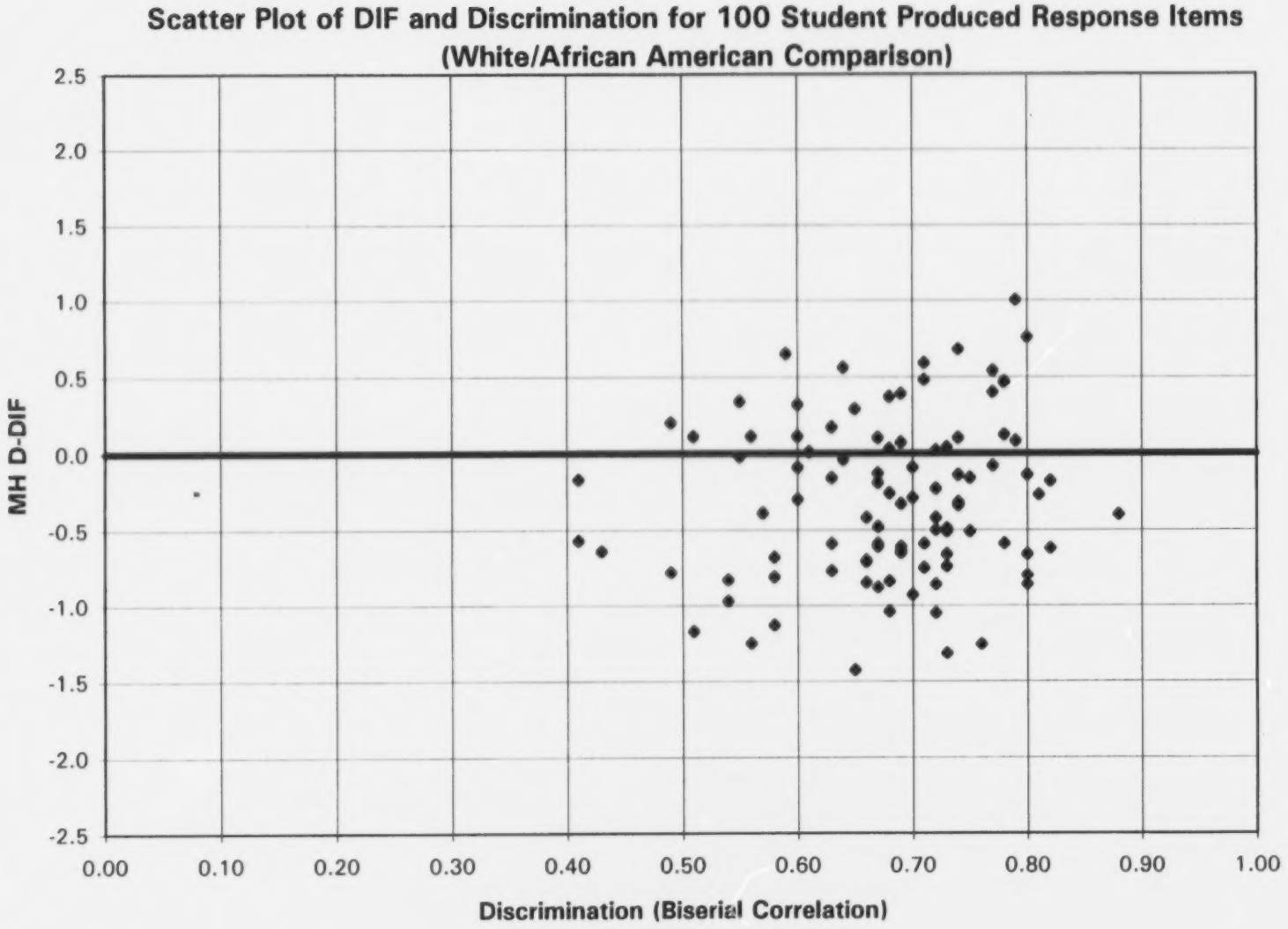
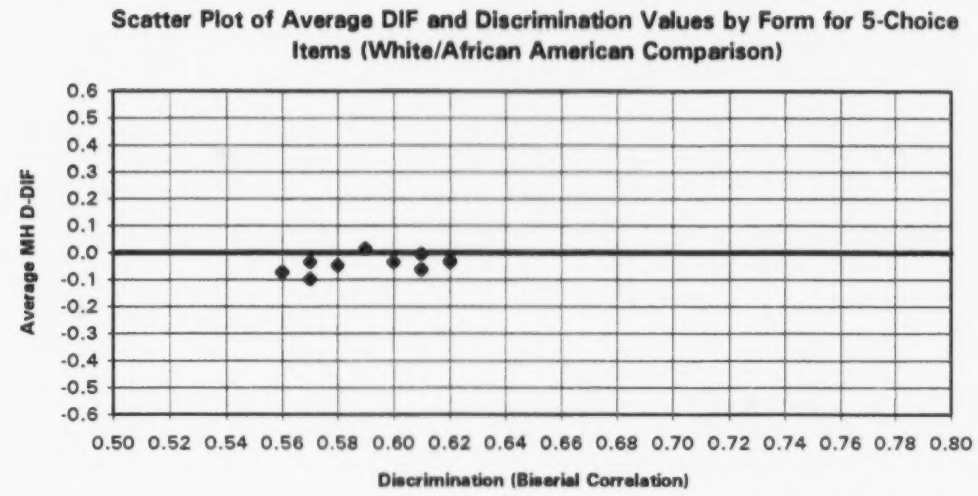
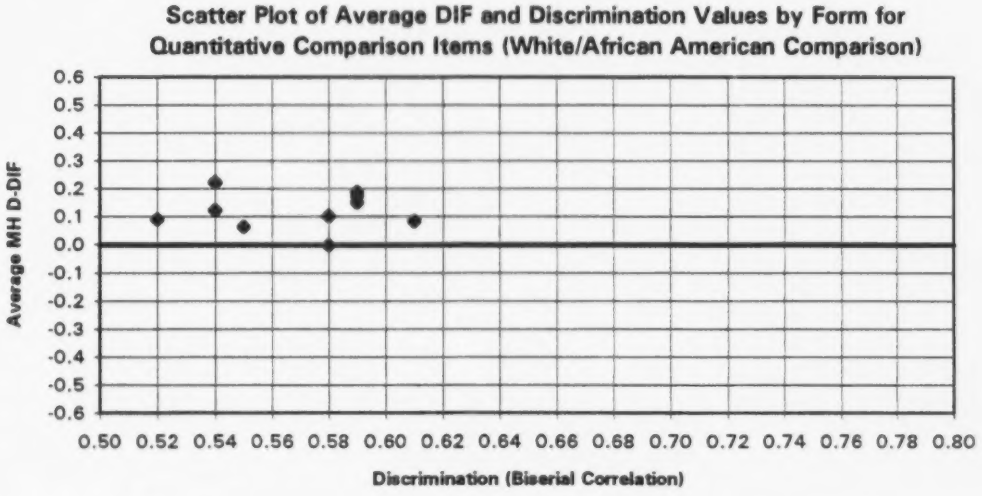
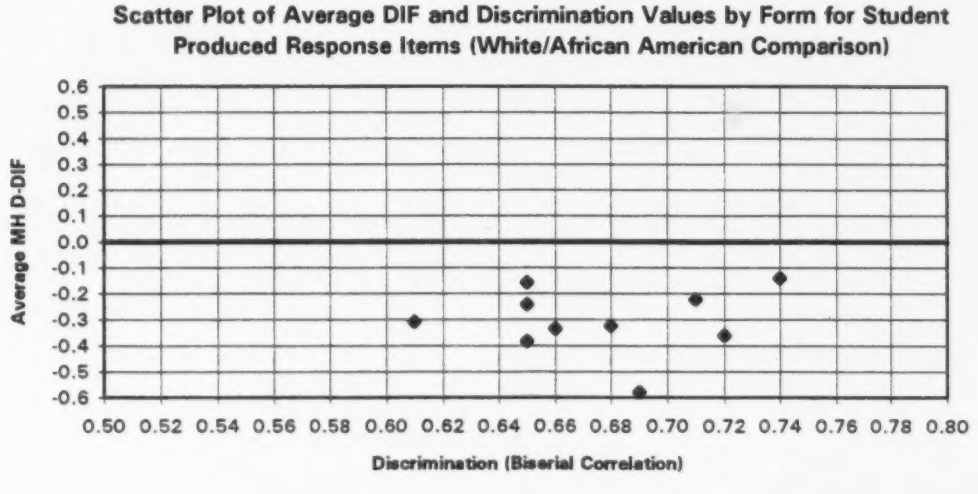

Scatter Plot of DIF and Discrimination for 100 Student Produced Response Items
(White/African American Comparison)
MH D-DIF
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Discrimination (Biserial Correlation)

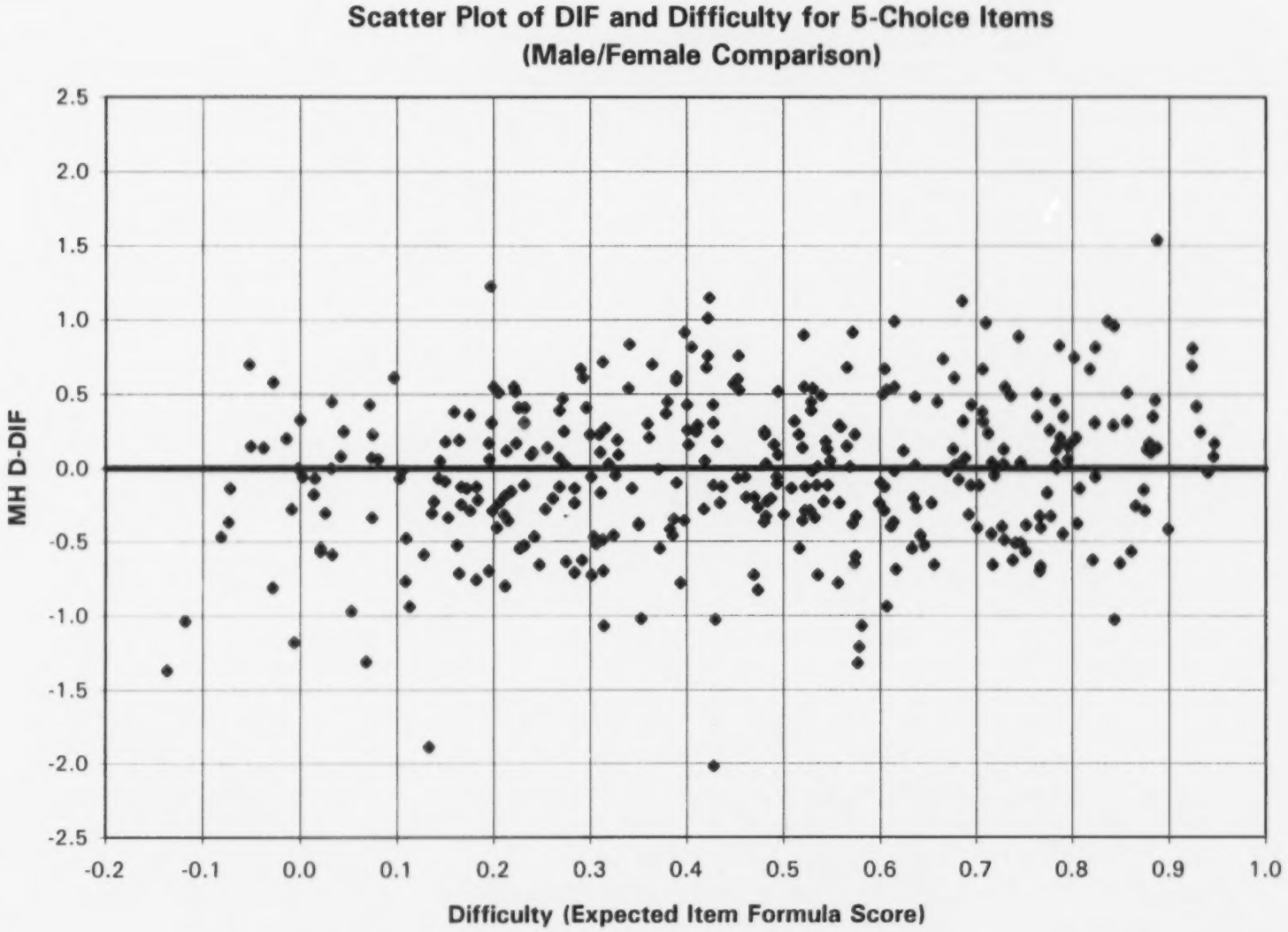
Figure 14


Scatter Plot of Average DIF and Discrimination Values by Form for 5-Choice Items (White/African American Comparison)
Average MH D-DIF
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
0.50 0.52 0.54 0.56 0.58 0.60 0.62 0.64 0.66 0.68 0.70 0.72 0.74 0.76 0.78 0.80
Discrimination (Biserial Correlation)

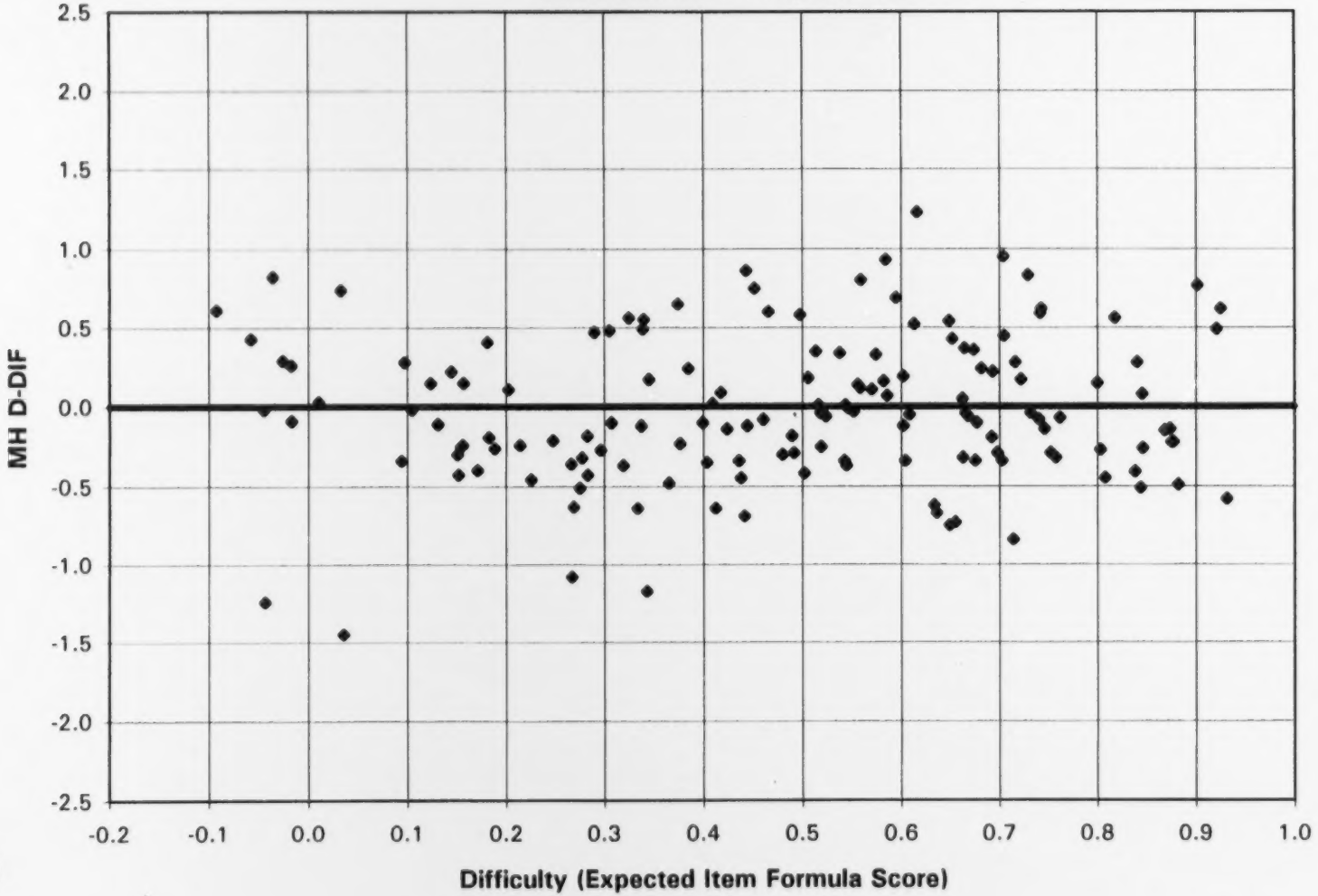

Scatter Plot of Average DIF and Discrimination Values by Form for Quantitative Comparison Items (White/African American Comparison)
Average MH D-DIF
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
0.50 0.52 0.54 0.56 0.58 0.60 0.62 0.64 0.66 0.68 0.70 0.72 0.74 0.76 0.78 0.80
Discrimination (Biserial Correlation)


Scatter Plot of Average DIF and Discrimination Values by Form for Student Produced Response Items (White/African American Comparison)
Average MH D-DIF
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
0.50 0.52 0.54 0.56 0.58 0.60 0.62 0.64 0.66 0.68 0.70 0.72 0.74 0.76 0.78 0.80
Discrimination (Biserial Correlation)

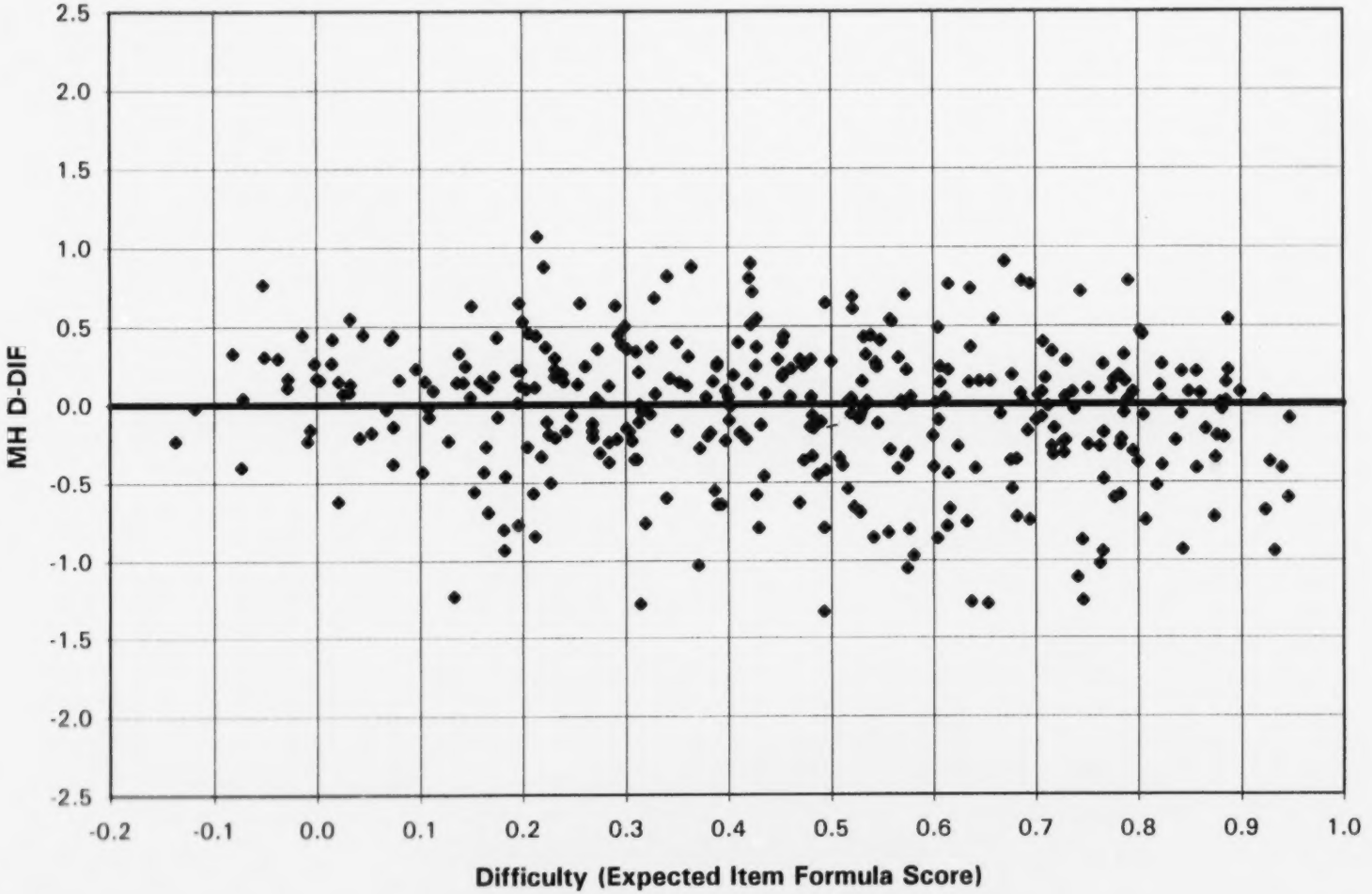
A.1


Scatter Plot of DIF and Difficulty for 5-Choice Items
(Male/Female Comparison)
MH D-DIF
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Difficulty (Expected Item Formula Score)

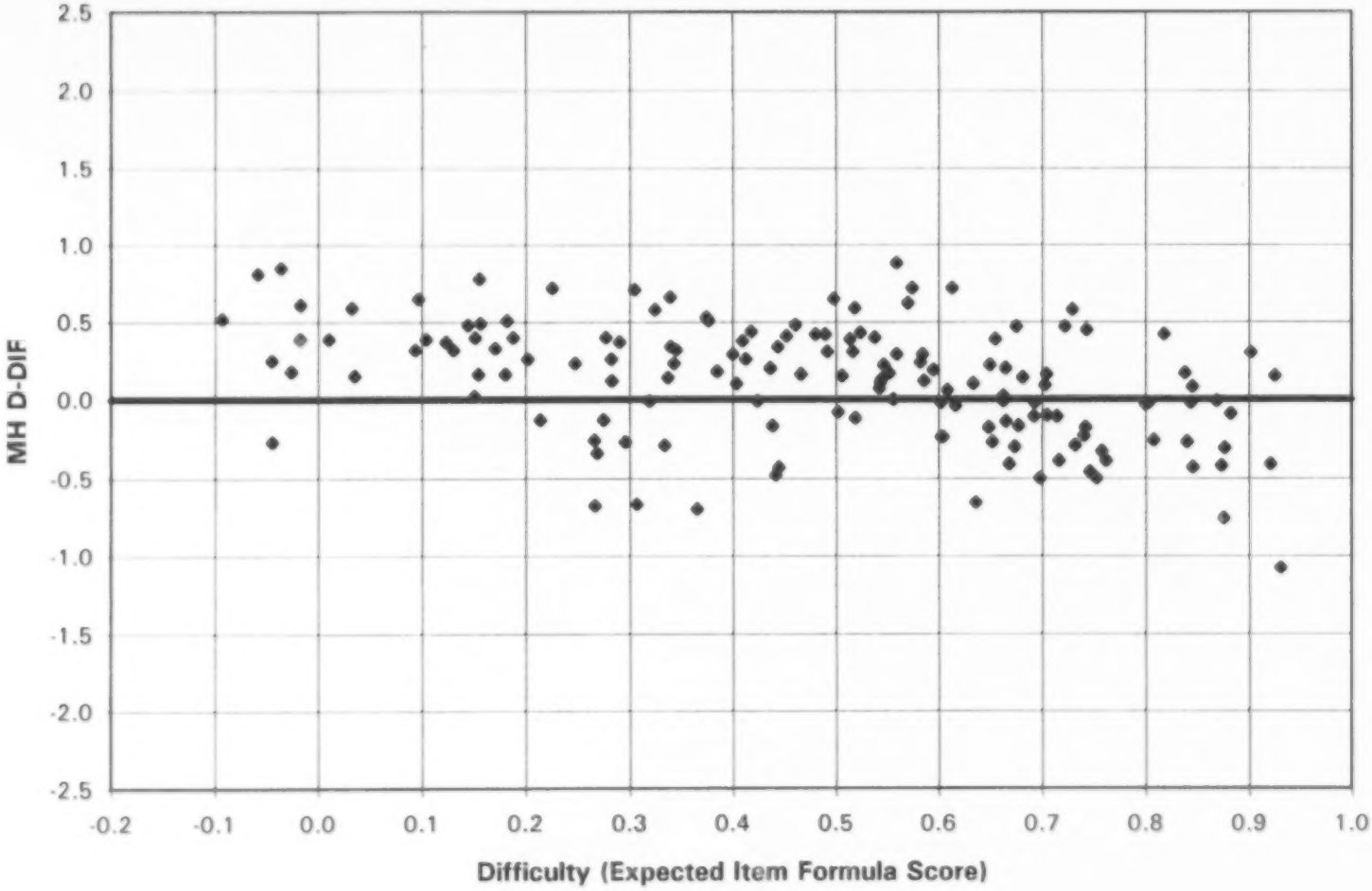
Scatter Plot of DIF and Difficulty for Quantitative Comparison Items (Male/Female Comparison)



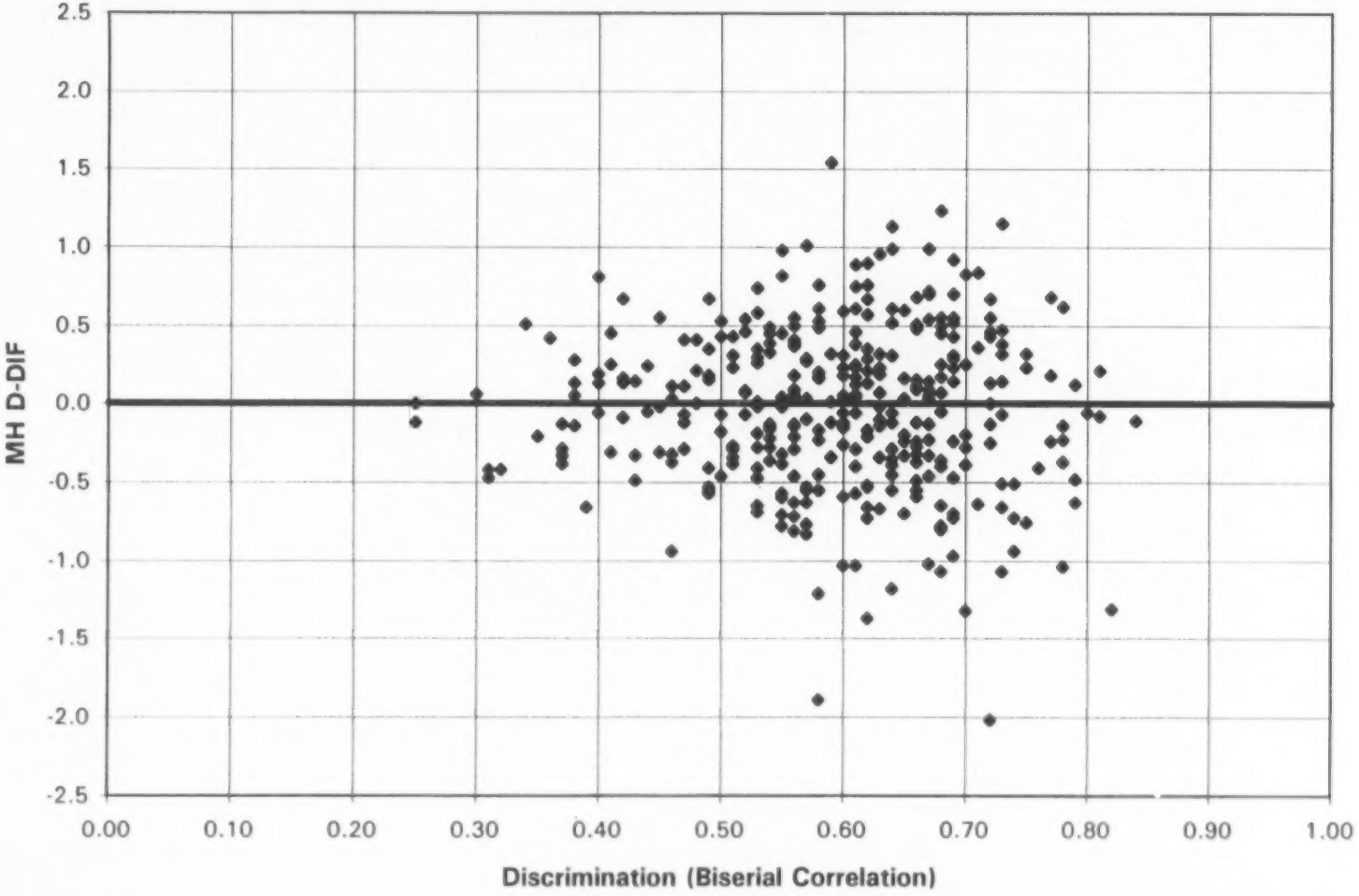
Scatter Plot of DIF and Difficulty for 5-Choice Items (White/African American Comparison)



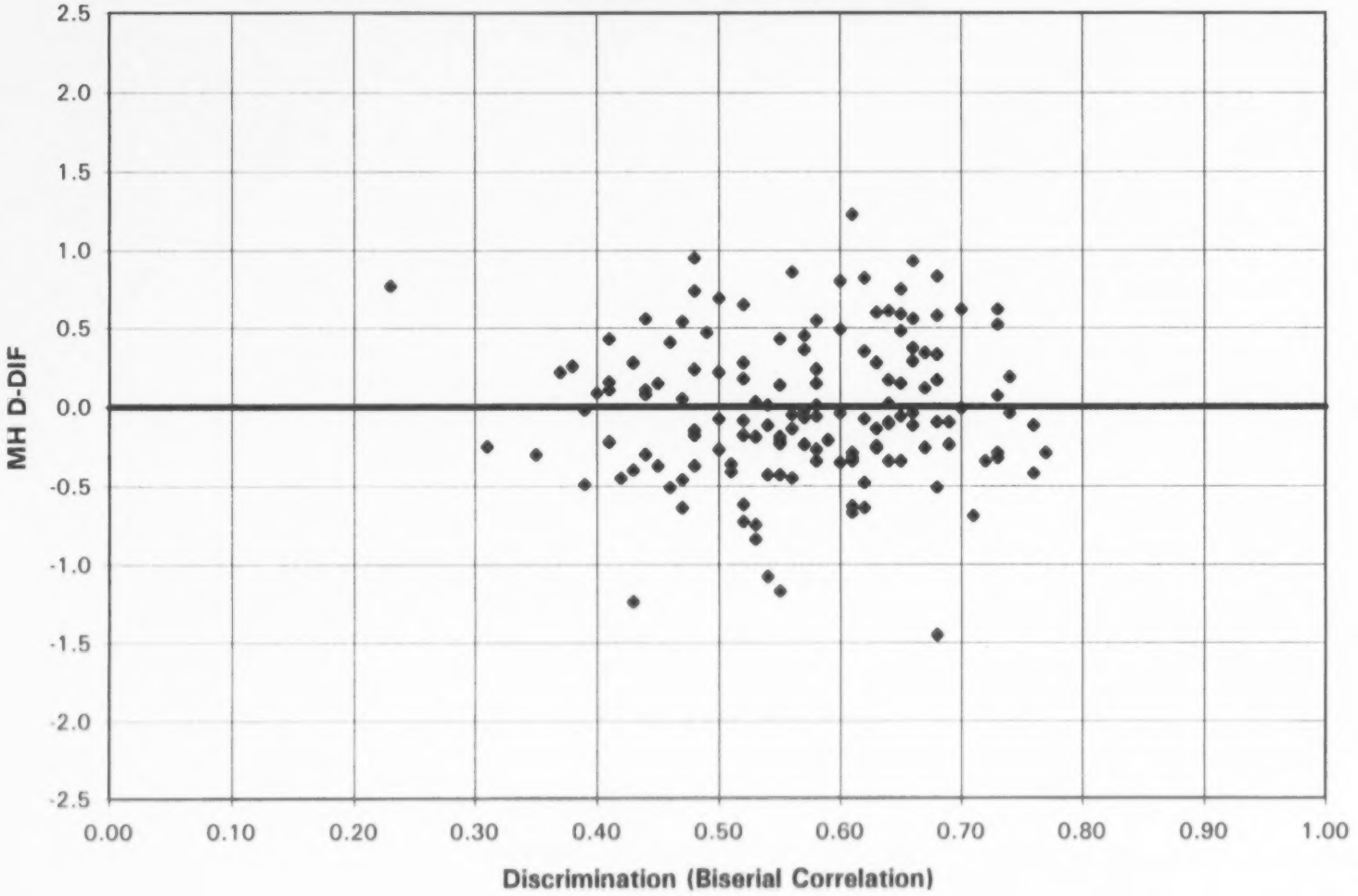
Scatter Plot of DIF and Difficulty for Quantitative Comparison Items (White/African American Comparison)



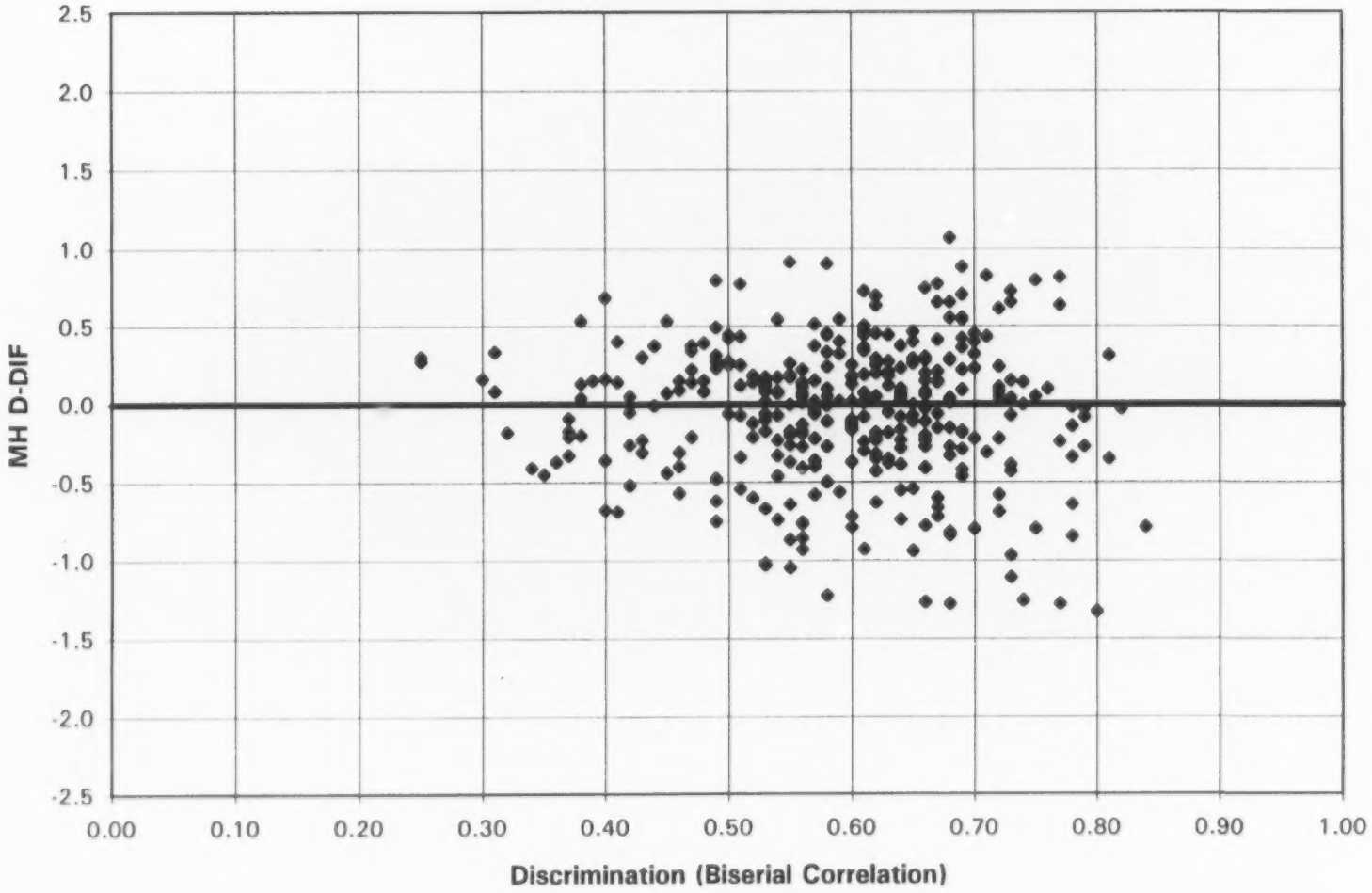
Scatter Plot of DIF and Discrimination for 5-Choice Items (Male/Female Comparison)



Scatter Plot of DIF and Discrimination for Quantitative Comparison Items (Male/Female Comparison)



Scatter Plot of DIF and Discrimination for 5-Choice Items (White/African American Comparison)



Scatter Plot of DIF and Discrimination for Quantitative Comparison Items (White/African American Comparison)